Apport de la modélisation et des simulations de dynamique moléculaire à la description de STAT5 comme cible pour moduler la signalisation oncogénique

Florent Langenfeld

UNIVERSITÉ PARIS-SUD

ÉCOLE DOCTORALE 418 :
DE CANCÉROLOGIE

Laboratoires :
Laboratoire de Biologie et Pharmacologie Appliquée, CNRS – ENS Cachan
Centre de Mathématique et de Leurs Applications, CNRS – ENS Cachan

THÈSE DE DOCTORAT

Discipline :
ASPECTS MOLÉCULAIRES ET CELLULAIRES DE LA BIOLOGIE

par

Florent LANGENFELD

Apport de la modélisation et des simulations de dynamique moléculaire à la description de STAT5 comme cible pour moduler la signalisation oncogénique

Date de soutenance : 05/06/2015

Composition du jury :

Directrice de thèse :	Luba Tchertanov	Directrice de recherche (CMLA UMR 8536)
Co-directeur de thèse :	Michel AROCK	Professeur (ENS Cachan)
Rapporteurs :	Alexandre DE BREVERN	Directeur de recherche (INSERM UMR_S 1134)
	Pascal BONNET	Professeur (Université d'Orléans)
Examinateurs :	Éric SOLARY	Professeur (Université Paris – Sud)
	Marc BAADEN	Directeur de recherche (CNRS UPR 9080)
	Francesco PIAZZA	Professeur (Université d'Orléans)

Résumé

STAT5 est une protéine de la signalisation cellulaire normale, ayant un rôle important dans la transformation, la survie et à la résistance aux inhibiteurs de tyrosine kinase des cellules tumorales. Son activation constitutive par phosphorylation est liée à la présence de protéines oncogéniques comme la protéine de fusion BCR/ABL1 (leucémie myéloïde chronique) ou de formes mutées de KIT (mastocytoses) notamment. L'inhibition pharmacologique de STAT5 constitue donc un enjeu thérapeutique majeur pour plusieurs pathologies malignes. Nous avons réalisé la première modélisation et les simulations de dynamique moléculaire des principales formes de STAT5 : la forme monomérique cytoplasmique phosphorylée ou non, et la forme dimérique phosphorylée et liée à l'ADN. Nous avons caractérisé les propriétés dynamiques et le réseau allostérique intramoléculaire des monomères de STAT5. Les résultats générés montrent des variations structurales et dynamiques liées aux variations de séquence primaire des isoformes de STAT5 et/ou à la présence du groupement phosphate. Deux poches à la surface des protéines ont également été caractérisées. Leur localisation à proximité de voies de communication allostériques suggère que ces poches pourraient constituer des sites de modulation des fonctions de STAT5. Nous avons également caractérisé les liaisons hydrogènes entre les monomères constituant les dimères de STAT5 et leur reconnaissance de l'ADN. En outre, nous avons identifié des résidus clés aux interfaces entre les entités moléculaires, nous permettant de mieux comprendre les effets de mutations de STAT5 observées en clinique.

Summary

STAT5 is a protein involved in normal cell signalling that is crucial for transformation, survival and resistance to tyrosine kinase inhibitors of tumour cells. The constitutive phosphorylation activates STAT5 and is related to oncogenic proteins like the hybrid protein BCR/ABL1 (chronic myeloid leukaemia) or mutated KIT receptor (mastocytosis). The pharmacologic inhibition of STAT5 is thus a major therapeutic concern in several malignant pathologies. We performed the first modelling and molecular dynamics simulations of the main cellular species of STAT5: the cytoplasmic phosphorylated or unphosphorylated monomer, and the phosphorylated dimer bound to DNA. We characterized the dynamical properties and the intramolecular allosteric network of the monomers. The generated results show structural and dynamic variations linked to the primary sequence changes between the two STAT5 isoforms and/or to the phosphate group. Two pockets were characterized at the surface of STAT5. Their location at close proximity of allosteric communication pathways suggests new putative inhibition sites to modulate STAT5 functions. We also described the hydrogen bonds network between the monomers of the dimeric species and the recognition of the DNA. We identified key residues at the interfaces, allowing us to better understand the effects of clinically relevant STAT5 mutations observed in malignancies.

Table des matières

Table des figures

Table des tableaux

Chapitre 1 : Introduction

PRÉAMBULE

On estime à environ 10^{14} le nombre de cellules chez un être humain, et à plus de 200 le type de cellules qui coexistent dans les tissus. Chaque type cellulaire peut exercer des fonctions extrêmement différentes, dépendantes à la fois de sa localisation tissulaire mais également de son environnement. Afin d'assurer le fonctionnement optimal de cet ensemble hétérogène, les cellules communiquent entre elles *via* des signaux qui peuvent être de nature différente (chimiques, électriques, …) et prendre différentes formes (*via* un contact direct - signalisation juxtacrine -, à de courte distance - signalisation paracrine -, ou à grande distance - signalisation endocrine -. Les différentes fonctions du corps humain sont ainsi régulées par un ensemble de signaux qui parcourent parfois des distances considérables afin de délivrer le signal adéquat à un groupe de cellules bien déterminé. Ce signal à grande échelle est ensuite transmis *in cellulo* par d'autres acteurs moléculaires qui vont soit prolonger la transmission du signal, soit effectuer une opération en réponse au signal qu'ils intègrent. Les messagers inter-tissulaires ou intercellulaires sont très souvent incapables, du fait de leur nature généralement polaire, de franchir la barrière qu'est la membrane cellulaire lipidique. Par conséquent, chaque cellule présente en surface des récepteurs transmembranaires qui filtrent les signaux afférents et assurent la transmission d'un signal adéquat du milieu extracellulaire vers le milieu intracellulaire. Puis, au niveau cytoplasmique, la transmission du signal transite de proche en proche par l'activation successive d'agents moléculaires généralement de nature protéique, jusqu'à la cible qui va transformer le signal en une réponse cellulaire. Le signal aura parcouru dans la cellule l'ensemble de ce qu'on appelle couramment une **voie de signalisation**, les protéines activées étant dès lors appelées **protéines de signalisation**.

Les protéines STATs (*Signal Transducer and Activator of Transcription*) font partie de ce groupe des protéines de signalisation comme leur nom l'indique, mais sont également les molécules effectrices du signal. Ainsi, ce sont des molécules primordiales dans la régulation d'un grand nombre de processus physiologiques. La transmission du message physiologique à l'échelle cellulaire s'accompagne de la transmission d'un message purement chimique à l'échelle moléculaire, sous la forme d'un groupement phosphate qui est transmis aux protéines STATs. Cet évènement est central et initie l'activité de ces protéines, qui consiste en l'activation de la transcription de gènes cibles qui leur procure la seconde partie de leur appellation. La transcription ciblée de gènes induit l'expression de protéines qui vont apporter la réponse cellulaire consécutive à l'activation d'une voie de signalisation. Cependant, d'autres aspects entrent en considération, comme l'intégration simultanée de plusieurs voies de signalisation pouvant avoir des effets opposés, ou l'interaction de plusieurs réseaux de protéines associés à des voies de signalisation différentes.

Particulièrement, dans certains contextes pathologiques, l'équilibre entre les différentes voies de signalisation est rompu par l'introduction d'un élément perturbateur qui favorise ou inhibe l'une ou l'autre des voies de signalisation. Ainsi, les protéines STAT5 sont parfois activées continuellement suite à un signal tronqué qui touche les protéines en amont de cette protéine

STAT. Dès lors, cette famille de protéine est apparue comme un acteur majeur de la pathogénie de plusieurs cancers hématologiques, mais également de plusieurs tumeurs solides. De nombreuses études ont montré le bénéfice qu'il y aurait à cibler ces protéines dans le cadre de stratégie anti-cancéreuses afin d'apporter de nouvelles approches thérapeutiques. Cependant, aucun composé spécifique et utilisable à des doses pharmacologique n'est disponible actuellement, et l'absence de données structurales portant sur STAT5 rend la recherche de tels composés difficile. Ces éléments sont présentés dans l'introduction de ce travail de thèse.

L'objectif des travaux de cette thèse est de générer les premières données structurales des différentes formes de STAT5, mais également d'étudier sa dynamique à l'échelle moléculaire et atomique afin de proposer des stratégies innovatrices d'inhibition/modulation basées sur de nouveaux sites d'inhibition. Pour cela, nous avons utilisé un ensemble de techniques de biologie computationnelle, présentées dans la seconde partie de l'introduction, afin de caractériser les éléments structuraux et dynamiques en relation avec la fonction de STAT5. Les détails des calculs et des méthodes utilisées et parfois développées au cours de ce travail sont ensuite décrites dans le second chapitre.

Dans un troisième et dernier temps, nous présentons les résultats extraits des données que nous avons générées ainsi que leur interprétation au regard des données expérimentales disponibles. Nous discutons également au cours de cette partie des éléments qui pourraient amener le développement de nouveaux composés innovants ciblant spécifiquement STAT5.

I. Les protéines STATs et STAT5 : Organisation, rôles physiologiques et implications pathologiques

A. La famille des protéines STATs

1. *Les STATs : gènes, organisation structurale et cycle d'activation – désactivation.*

La famille de protéine STAT (*Signal Transducer and Activator of Transcription*) est composée chez les mammifères de 7 protéines partageant un mode de fonctionnement ainsi qu'une organisation structurale et des propriétés similaires. Les protéines STAT assurent la transmission d'un signal cellulaire depuis un récepteur, qui peut être membranaire ou cytoplasmique, jusqu'au noyau et à l'ADN chromatinien.

Ces protéines ont été caractérisées au cours des années 90, et les gènes codant pour les STATs sont situés sur les chromosomes 2 (STAT1[1] et STAT4[2,3]), 12 (STAT2[1] et STAT6[4,5]) ou 17 (STAT3[6], STAT5a[7,8] et STAT5b[8]). L'analyse des gènes murins correspondants semblent indiquer que ces gènes dérivent d'un gène initial commun qui aurait subi une duplication initiale sur le même gène. Deux duplications en tandem successives auraient suivi, produisant ainsi 3 copies du gène STAT primordial sur 3 chromosomes différents. Ces 3 gènes auraient ensuite évolués, la dernière duplication donnant naissance aux 2 gènes codant pour STAT5a et STAT5b[9]. Ce schéma explique la très forte similarité de STAT5a et STAT5b (en termes de séquence d'acide aminé) par comparaison aux autres STATs (*cf.* Tableau 1).

Tableau 1: **Similarité des séquences primaires antre chaque protéine STAT (exprimée en pourcentage).**

	STAT1	STAT2	STAT3	STAT4	STAT5a	STAT5b
STAT6	21,73	20,43	20,26	22,99	33,25	33,16
STAT5b	26,53	24,02	26,62	28,48	92,88	
STAT5a	28,27	23,55	27,27	28,88		
STAT4	52,54	39,71	46,93			
STAT3	51,87	35,06				
STAT2	40,80					

Malgré la faible similarité de séquence de certaines protéines STATs (*cf.* Annexe A.), l'organisation structurale des différents domaines est commune à l'ensemble de la famille. Elle consiste en : (i) un domaine situé à l'extrémité N-terminale des STATs (*N-terminal Domain* : ND) et composé majoritairement d'hélices α, (ii) d'un domaine de type « Coiled-Coil » (*Coiled Coil Domain* : CCD) composé de quatre hélices α, (iii) d'un domaine de liaison à l'ADN (*DNA Binding Domain* : DBD) principalement caractérisé par des feuillets β, (iv) d'un domaine intermédiaire

(*Linker Domain* : LD) α-hélical, (v) un domaine « *Src Homology 2* » SH2 mixte composé d'hélices α et de feuillets β, et (vi) un domaine C-terminal de transactivation (*TransActivation Domain* : TAD). La structuration de ce dernier domaine reste méconnue, et le TAD est vraisemblablement non-structuré mais capable de former au moins une hélice α dans certaines conditions[10].

Les protéines STATs partagent un cycle d'action commun consistant en une activation cytosolique, un import nucléaire, une activité nucléaire de promotion de transcription, un export nucléaire et une désactivation. A la fin de ce cycle, les protéines rejoignent le pool cytoplasmique de STATs latents et activables. Les ligands activateurs de STAT diffèrent largement. Un grand nombre de cytokines (interférons, facteurs de croissance, interleukines, …) pourraient être citées, alors qu'un nombre plus réduit de protéines assure l'activation des STATs. Les protéines de la famille JAK (*Janus Kinases*) ont été les premières dont l'implication dans l'activation des STATs a été prouvée[11], permettant d'établir le schéma d'activation suivant : (i) la liaison d'une cytokine à son récepteur membranaire entraîne son oligomérisation puis l'activation des JAKs associées au récepteur, (ii) JAK trans-phosphoryle le domaine intracellulaire du récepteur, créant des sites d'amarrage pour les protéines STATs latentes, (iii) le recrutement des STATs se fait *via* le domaine SH2 et un résidu tyrosine (Tyr, Y) spécifique est phosphorylé en position C-terminale du domaine SH2, (iv) l'homo- ou hétéro-dimérisation des STATs est réalisée suite à cette phosphorylation, (v) le dimère est relâché, se transloque dans le noyau et se fixe sur des portions d'ADN spécifiques en amont de gènes dont la transcription est ensuite activée. Cet enchainement d'évènements cellulaires et moléculaires constitue la voie de signalisation canonique JAK/STAT (*cf.* Figure 1). Cependant, de récentes recherches sont venues compléter ce paradigme et apportent des éléments nouveaux quant à l'activation, la désactivation et la régulation des STATs.

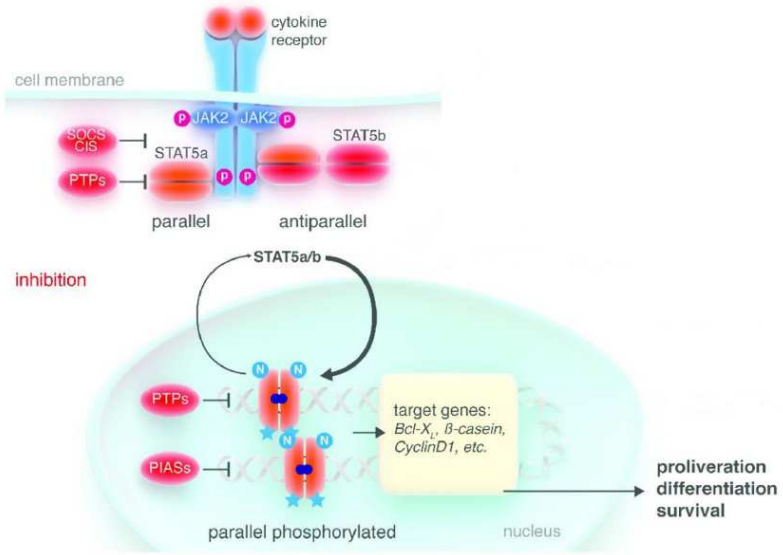


Figure 1: **La voie de signalisation JAK-STAT canonique.** Figure reproduite avec la permission des auteurs[12].

2. *L'activation des STATs par phosphorylation*

Il a été montré que l'évènement critique de l'activation des protéines STATs est la phosphorylation d'un résidu Tyrosine, en position C-terminale du domaine SH2[13–15]. Cette étape est cruciale car elle permet la dimérisation par l'interaction du résidu phosphotyrosine d'un monomère avec le domaine SH2 de l'autre monomère, créant un double point d'ancrage. Par la suite, de nombreuses études ont porté sur les responsables de cette activation.

Le premier activateur de STATs à avoir été isolé est un membre de la famille des JAKs, TYK2[16] (*Tyrosine Kinase 2*), et les autres membres de la famille des JAKs ont également été impliqués dans l'activation des STATs[17] peu après. De nombreuses autres familles d'activateurs sont venues compléter le paradigme initial JAK-STAT et impliquent de nouvelles voies de signalisation capables d'activer l'une des STATs. Des récepteurs à activité tyrosine kinase intrinsèque tels que les récepteurs à l'EGF (*Epidermal Growth Factor*)[18], au PDGF (*Platelet-derived Growth Factor*)[19] ou au FGF (*Fibroblast Growth Factor*) [20] sont ainsi capables d'activer directement, ou indirectement, les STATs. Certains récepteurs couplés aux protéines G (RCPG) ou certaines tyrosines kinases cytoplasmiques sont également capables d'activer une protéine STAT : Ferrand et collaborateurs ont par exemple montré que JAK2/STAT3 est activée par un RCPG, le RCCK2 (récepteur de la cholécystokinine-2), dans le pancréas de souris[21], Wong et Fish ont également démontré que les chimiokines RANTES (*Regulated on activation, normal T cell expressed and secreted*) et MIP-1α (*Macrophage inflammatory protein – 1 α*) régulent les fonction des cellules T et que cette régulation est médiée grâce à l'activation de STAT1 et STAT3 par des RCPGs (CCR1, CCR4 et CCR5 – *chemokines receptor1, 4, 5*)[22]. De manière similaire, Danial et collaborateurs ont montré dans des cellules BAF/3 transformées, que la protéine virale pro-oncogène v-Abl se lie à JAK1 pour activer constitutivement STAT1, STAT3 et STAT5[23], alors que l'isoforme M2 de la pyruvate kinase (PKM2) a récemment été décrite comme phosphorylant STAT3 au niveau nucléaire[24].

Ainsi, si la voie de signalisation JAK-STAT reste celle la plus souvent activée, de nombreuses autres protéines sont impliquées dans l'activation des STATs et viennent enrichir le réseau de signalisation des protéines STATs. Cependant, la phosphorylation d'un seul résidu tyrosine n'explique pas à elle seule la variation d'activité parfois observée au sein des cellules. D'autres phénomènes physiologiques sont ainsi à l'œuvre et participent à la modulation de l'activité des STATs au niveau cellulaire.

3. *Déphosphorylation et formes tronquées des STATs (régulation négative de l'activité des STATs)*

Si l'activation des protéines STATs est étroitement liée à un évènement de phosphorylation, la déphosphorylation du résidu de phosphotyrosine constitue le premier élément de régulation des protéines STATs et de l'expression des gènes dont la transcription est régulée par les STATs, donc pour empêcher la progression/transmission de signaux

oncogéniques. Les éléments cellulaires responsables de la déphosphorylation sont donc des éléments cruciaux du cycle de phosphorylation/déphosphorylation, au même titre que les activateurs de STATs tels que les récepteurs tyrosine kinases. Trois familles de protéines assurent la régulation négative des protéines STATs (*cf.* Figure 1): les PTPs (phosphotyrosines phosphatases), les SOCS (*Suppressor of cytokine signaling*) et les PIAS (*Protein inhibitor of activated STAT*).

Les PTPs constituent une famille de protéines composée de plus de 100 membres, transmembranaires, cytoplasmiques ou nucléaires, qui partagent certains motifs structuraux en lien avec leur activité phosphatase. Ce groupe de protéines peut à la fois agir directement par déphosphorylation des STATs ou indirectement par déphosphorylation des activateurs des STATs (protéines JAKs, notamment). La nature des PTPs impliquées dans la régulation des STATs dépend essentiellement de l'isoformes de STAT considérée et du type cellulaire. Par exemple, il a été montré que STAT6 est déphosphorylé par SHP-1 (*SH2-domain containing phosphatase*) dans les précurseurs hématopoïétiques[25] et les cellules T[26], TCPTP (*T cell protein tyrosine phosphatase*) dans le noyau de lymphocytes B activés[27], PTP1B (*phosphotyrosine phosphatase* 1B) dans des fibroblastes[28] et PTP-BL (*phosphotyrosine phosphatase BL*)[29]. Mais SHP-1 est impliqué également dans la déphosphorylation de STAT3/5a/5b/6, TCPTP dans la régulation de STAT1, PTP1B n'agit que sur STAT6 alors que PTP-BL peut également déphosphoryler STAT4 en plus de STAT6. Les membres de la famille des PTPs possèdent ainsi une spécificité limitée à une ou plusieurs STATs, à l'exception de STAT2 qui n'a jamais été rapportée dans la littérature comme cible d'une PTP (*cf.* Tableau 2, adapté de [30]). Cependant, certaines PTPs agissant sur STAT1 pourraient également réguler négativement STAT2.

Tableau 2 : Les protéines STATs et leur régulateurs PTPs.

STAT	PTPs régulatrices
STAT1	TCPTP, SHP-2
STAT2	
STAT3	TCPTP, SHP-2, PTPRD, PTPRT, SHP-1
STAT4	PTP-BL
STAT5a/b	VHR, SHP-2, SHP-1, ICA-512
STAT6	SHP-1, TCPTP, PTP1B, PTP-BL

La famille des protéines SOCS est composée de huit membres, SOCS1-7 et CIS (*Cytokine-induced SH2 containing protein*). Les SOCS sont des éléments permettant le rétrocontrôle négatif des voies de signalisation des cytokines ; elles vont donc impacter l'activation des familles de protéines impliquées dans la signalisation des cytokines, dont les STATs, par différents modes d'action: (i) les SOCS peuvent induire un mécanisme de compétition au niveau du recrutement des STATs par les récepteurs à cytokine ; (ii) les SOCS bloquent les résidus tyrosine du domaine d'auto-activation des JAKs, entrainant l'inhibition de ces protéines ; (iii) la dégradation de protéines cibles *via* le protéasome est également possible. Parmi cette famille,

CIS et SOCS1-3 sont les protéines les plus étudiées, et de fait le plus souvent reliées à la régulation des STATs. CIS est notamment reliée à la régulation de STAT5 au cours de la différenciation érythroïde[31], SOCS-1 est associée principalement à la régulation de STAT1 et STAT3 *via* l'inhibition des protéines JAKs[32], SOCS-2 régule la phosphorylation de STAT5[33] en se fixant notamment sur les récepteurs à l'hormone de croissance GH (*Growth Hormone*)[34], SOCS-3 régule également la phosphorylation de plusieurs STATs (comme SOCS-1) mais en modulant l'activité des JAKs en présence de récepteurs uniquement (contrairement à SOCS-1)[35]. Les autres SOCS sont moins bien connues car peu étudiées, mais il a été montré par exemple que l'expression constitutive de SOCS-5 inhibe l'activation de STAT6 par IL-4 (Interleukine 4)[36].

Une troisième famille de protéines est primordiale pour la régulation de l'activation des STATs, il s'agit des PIAS (*protein inhibitors of activated STAT*). Cette famille de protéines est composée de 5 membres qui sont exprimés constitutivement, contrairement aux membres de la famille des SOCS[37]. Leur interaction avec les protéines STAT sont néanmoins sujette à l'activation préalable de ces dernières[38]. Le mode d'action des protéines PIAS est différent en fonction de la PIAS considérée : elle peut soit bloquer les sites de liaison à l'ADN, soit promouvoir la SUMOylation des STATs[39,40]. D'autres mécanismes spécifiques d'un membre de la cette famille ont également été décrits[41,42].

Si le paysage global de l'inactivation des STATs par les familles de protéine SOCS, PIAS et PTPs reste incomplet, des phénomènes d'épissage alternatif ou de protéolyse post-traductionnel ont été notés, notamment au niveau du domaine de transactivation (TAD), et jouent un rôle dans la régulation de l'activité des STAT. Les formes non-tronquées des protéines sont dénommées isoformes α, tandis que les formes raccourcies sont appelées β, γ ou δ. Toutes les protéines STATs n'ont pas été répertoriées comme produisant des formes courtes. Ainsi, STAT1, STAT3, STAT4, STAT5a et STAT5b ont des formes β tronquées au niveau de l'extrémité C-terminale obtenue par épissage alternatif[43–46]. STAT6 présente 3 variants d'ADN complémentaire nommés STAT6a, STAT6b et STAT6c[47], présentant une délétion N-terminale (STAT6b) ou une délétion au sein du domaine SH2 (STAT6c). Les formes β présentent la même capacité à être phosphorylées sur le résidu tyrosine critique, à former des dimères et à lier l'ADN. Ainsi, les formes β de STAT1[15], STAT3[48], STAT5[46] et STAT6c[47] constituent des dominants négatifs des formes complètes. Les deux formes α et β des protéines STAT1, STAT3 et STAT4 présentent des activités distinctes caractérisées par l'activation de jeu de gènes différents[43,49,50]. Le clivage protéolytique de l'extrémité C-terminale produit également différentes formes de STAT3β, γ ou δ, STAT5β ou STAT6β[51–61]. L'origine de ces protéines tronquées est diverse : les différentes formes de STAT5β sont principalement retrouvées dans les cellules progéniteurs myéloïdes ou les cellules de patients atteints de leucémie aigüe myéloïde (LAM)[51,52,56,57,61], STAT6β dans les cellules mastocytaires de la moelle osseuse[59,60], STAT3β dans les cellules de patients avec LAM[61], STAT3γ dans les neutrophiles[53] et STAT3γ dans les formes très indifférenciées de granulocytes[55]. L'ensemble de ces formes permet la régulation fine des gènes dont l'expression est modulée par les protéines STATs.

4. Les autres sites de phosphorylation des STATs et les modifications post-traductionnelles

Suite à la découverte de cette nouvelle famille de protéines, il est rapidement apparu que les STATs possèdent des sites de phosphorylation additionnels et distincts du site de phosphorylation de la tyrosine. Un ou plusieurs sites de phosphorylation ont été découverts en fonction de l'isoforme considérée, principalement dans le domaine C-terminal des protéines de la famille STAT, le domaine de Trans-Activation (TAD), mais aussi dans d'autres domaines (ND et CCD). La plupart des protéines STATs des vertébrés possèdent un site de phosphorylation conservé contenant un résidu sérine, et localisé dans le TAD. Ce site est constitué d'un motif autour d'un résidu sérine conservé, de séquence P(M)SP (ou proche) vers la position 725 : un motif PMSP chez STAT1 et STAT3 (S727) ainsi que STAT4 (S721), un motif PSP chez STAT5a (S725) et STAT5b (S730), ou un motif SSPD chez STAT6 (S756). Une seconde sérine peut être présente, comme pour STAT1 (S708)[62] ou STAT5a (S779)[63]. D'autres sites de phosphorylation sur sérine sont également présents : un troisième site de phosphorylation est retrouvé sur STAT5a, au niveau du domaine N-terminal, en position S127/S128[64] alors que la sérine 287 (CCD) de STAT2 peut également être phosphorylée[65]. Les effets de ces différentes phosphorylations varient d'une isoforme à l'autre, ainsi que du site de phosphorylation considéré. Si les cytokines induisant la phosphorylation du résidu tyrosine critique sont également responsables de la phosphorylation des résidus de sérine dans le domaine TAD, la phosphorylation des résidus de sérine par d'autres ligands n'entraîne pas forcément de phosphorylation de la tyrosine, permettant une flexibilité de la modulation de l'activité des STATs *via* l'association à des protéines partenaires, association que serait dépendante de la phosphorylation des résidus de sérine.

Des modifications post-traductionnelles d'un autre type que la phosphorylation peuvent également survenir, sous le contrôle d'autres ligands. L'analyse de la littérature existante concernant ces modifications fait état plusieurs types de modifications, dans différents domaines, répertoriés ci-dessous.

L'acétylation réversible de résidus lysine est retrouvée dans toutes les STATs humaines, même si le nombre et la position des sites varient fortement d'une isoforme à une autre et continue à évoluer au fil des découvertes (*cf.* Figure 2). Ainsi, et contrairement à la phosphorylation du résidu tyrosine, les résidus de lysine le plus souvent acétylés diffèrent très largement en fonction de l'inducteur de l'acétylation, et induisent des effets variés en fonction de la protéine STAT ciblée et du résidu impliqué. Par exemple, l'acétylation de la lysine K685 de STAT3 joue un rôle critique dans la capacité de STAT3 à former des dimères stables et donc dans son activité[66]. *A contrario*, l'acétylation de la lysine 679 de STAT1, correspondant à K685 chez STAT3, ne semble pas avoir d'impact sur la phosphorylation de STAT1[67].

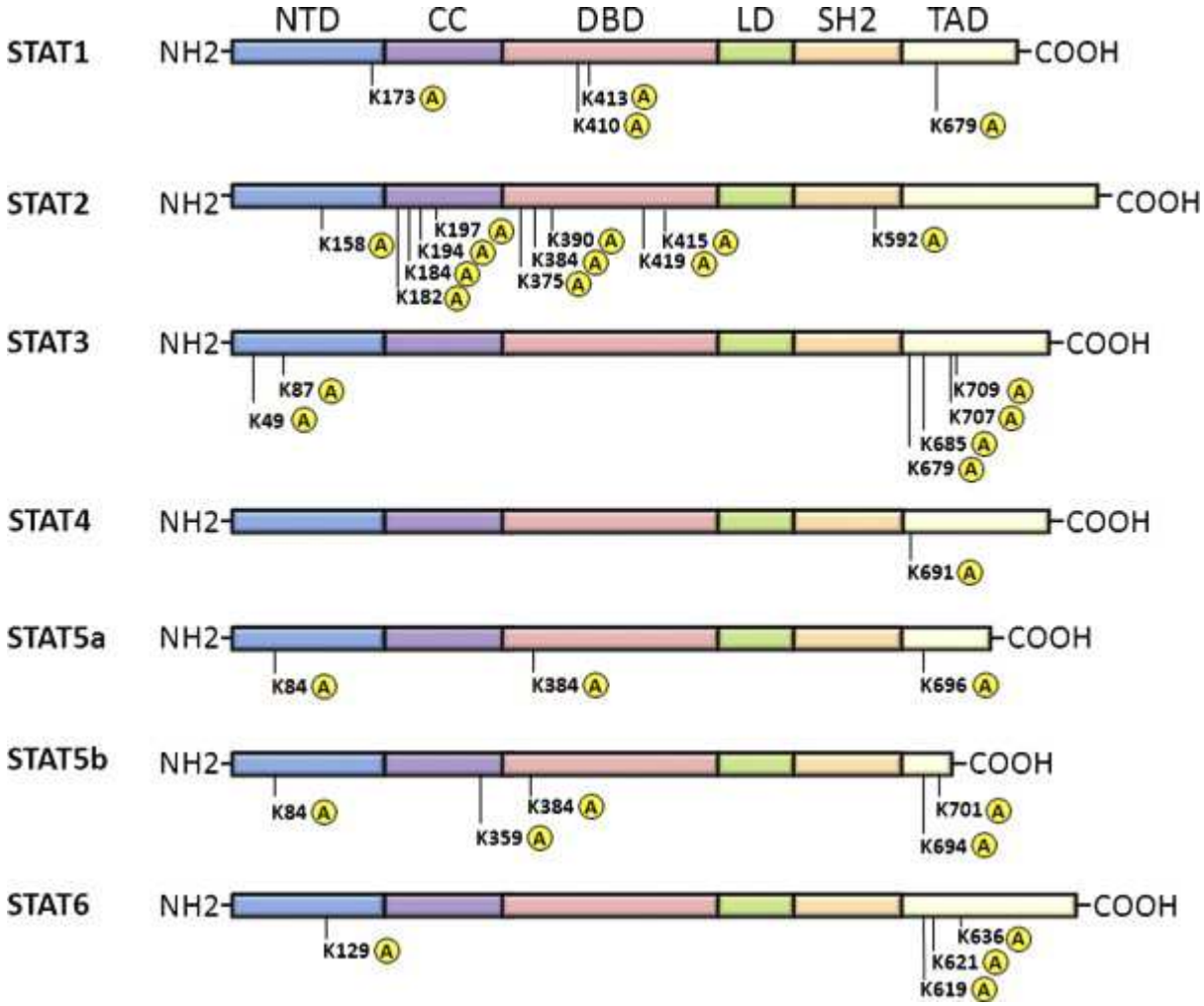


Figure 2: **Sites d'acétylation des STATs.** Figure reproduite avec la permission des auteurs[68].

La méthylation est un autre mode de régulation des protéines STATs, bien que moins répandu. La première mention de méthylation d'un résidu date de 2001, au cours d'une étude portant sur la modulation de l'activité de STAT1[69]. La méthylation de l'arginine 31 par la *Protein arginine N-methyltransferase 1* (PRMT1) a été montrée comme nécessaire à la promotion de la transcription induite par interféron α/β, les auteurs formulant l'hypothèse que la présence du groupement méthyl sur l'arginine 31 empêche l'interaction de STAT1 avec la protéine PIAS1 (*Protein inhibitor of activated STATs 1*), inhibant la déphosphorylation du résidu de tyrosine critique. La méthylation de résidus arginines a également été rapportée par la suite pour STAT6[70,71] et STAT3[72]. Des travaux publiés en 2004-2005 ont ensuite présenté des résultats contradictoires[73,74] et un échange avec les auteurs originaux[75] s'est engagé qui a débouché sur une absence de conclusion définitive et sur la nécessité de faire des études complémentaires. Depuis, de nouvelles études ont mis en avant la méthylation de l'arginine 31 de STAT3 [76], mais ont également évoqué du résidus K140[77], qui peut être diméthylé, et du résidu K180[78]. À l'image de l'acétylation des STATs, le rôle des résidus méthylés ou diméthylés varie : si la méthylation de K180 promeut la phosphorylation du résidu Y705 chez STAT3, la diméthylation de K140 va inactiver STAT3.

L'ubiquitination est une modification post-traductionnelle très générique, une ou plusieurs molécules d'ubiquitine venant se greffer généralement sur un résidu lysine et orientant la protéine ainsi ubiquitinée vers la voie de dégradation protéasome-dépendante. L'ubiquitination des STATs a été observée pour STAT1[79], STAT4, STAT5 et STAT6[10], STAT2[80] et STAT3[81], apportant un autre mécanisme de régulation de l'activité des protéines STATs.

De manière similaire, la SUMOylation est une modification post-traductionnelle qui consiste en la liaison covalente de protéines SUMO (*Small ubiquitin-like Modifier*) sur un résidu lysine, afin de réguler l'activité biochimique, la localisation cellulaire ou la stabilisation des protéines SUMOylées. Seule la protéine STAT1 a été décrite comme pouvant être SUMOylée sur le résidu lysine 703 par PIASx-α (*Protein inhibitors of activated STAT*)[82]. Étant donné sa proximité avec le site de tyrosine-phosphorylation (Y701 pour STAT1), les auteurs ont étudié l'impact de la SUMOylation sur l'activité et ont observé un effet négatif sur la transcription des gènes régulés par STAT1[83], en interférant avec la capacité de STAT1 à former des dimères stables[84].

Le processus d'ISGylation est également semblable à l'ubiquitination et à la SUMOylation puisqu'il consiste en l'adjonction d'ISG15 (la protéine associée à l'expression du gène *Interferon stimulated gene 15*) sur un résidu de lysine. STAT1 est le seul membre connu de la famille des STATs ciblé par une ISGylation[85] qui prolonge dans le temps la phosphorylation du résidu tyrosine critique pour l'activité de STAT1[86]. L'ISGylation n'est cependant pas requise à l'activité de STAT1[87].

Enfin, une équipe a identifié en 2004 des sites de glycosylation sur les formes nucléaires de STAT5a[88]. Plus précisément, certains résidus de sérine et thréonine de la fraction nucléaire de STAT5a sont conjugués à des molécules de N-acetylglucosamine après phosphorylation du résidu de tyrosine 694. L'un des sites de glycosylation a été identifié comme le résidu T92, dans le domaine N-terminal, et la mutation de ce résidu en alanine a induit la forte réduction de la glycosylation de STAT5a, mais a également montré que d'autres sites de glycosylation existaient. La glycosylation de ce résidu permet l'interaction STAT5a / CBP (*CREB-binding protein*), un co-activateur de la transcription qui potentialise l'activité de plusieurs familles de facteurs de transcription. Ainsi, la formation de ce complexe STAT5a / CBP, abolie par la mutation T92A, va activer la transcription des gènes induits par les voies de signalisation impliquant STAT5a, transcription absente dans les cellules mutées. La glycosylation de STAT1/3/5b/6 a également été montrée[88].

L'ensemble de ces modifications post-traductionnelles offrent ainsi des voies de régulation cellulaires qui permettent la modulation fine de l'activité de la famille des STATs, en sus du processus de phosphorylation/déphosphorylation. On peut également apercevoir dans la régulation des STATs d'importantes divergences, certains mécanismes étant spécifiques à une STAT donnée (ISGylation de STAT1, sites d'acétylation, *etc.*), montrant ainsi que toutes les STATs doivent être considérées indépendamment et non vues comme une famille parfaitement homogène en dépit du fait que le mécanisme d'activation le plus important – la phosphorylation – est commun à toutes les protéines STAT.

5. L'import/export nucléaire

Les chromosomes des cellules eucaryotes se trouvent à l'intérieur du noyau, où différents évènements de la régulation cellulaire ont lieu (réplication de l'ADN, transcription, …). Le noyau est séparé du cytoplasme par la membrane nucléaire qui régule l'accès au noyau des protéines et d'autres molécules (ARNs, …)[89]. Les protéines STATs sont activées (c'est à dire phosphorylées) au sein du cytoplasme, et doivent franchir cette barrière pour se fixer sur les sites promoteurs des gènes qu'elles régulent[90,91]. Le passage du milieu cytoplasmique vers l'intérieur du noyau constitue donc une étape importante dans l'activité des protéines STATs. Les mécanismes d'import et d'export nucléaire des STATs constituent donc un levier important dans la régulation du cycle de cette famille de protéine. Cependant, toutes les STATs ne sont pas prises en charge par les mêmes partenaires protéiques lors du passage de la barrière nucléaire.

La protéine STAT1 est principalement retrouvée dans le cytoplasme mais s'accumule de manière transitoire dans le noyau dans les minutes qui suivent une activation par l'interféron γ, avant de disparaître du noyau en quelques heures[92]. D'autres résultats indiquent que la disparition de STAT1 nucléaire est liée à sa déphosphorylation plutôt qu'à sa dégradation[93]. La phosphatase nucléaire TC45 a ainsi été montrée comme étant responsable de la déphosphorylation de STAT1[94], alors que l'exportine Crm1 est impliquée dans l'export nucléaire[95]. Crm1 se lie à un motif protéique consensus qui porte un signal d'export nucléaire (*nuclear export signal, NES*) situé dans le domaine de liaison à l'ADN (DBD)[95]. L'import nucléaire est principalement lié à la fois à la phosphorylation de STAT1 et sa liaison à l'importine α5[96]: la forme non-phosphorylée de STAT1 n'est en effet pas importée par l'importine α5. Cependant, des formes non-phosphorylées de STAT1 ont été détectées dans le noyau[97,98]. Le mécanisme d'import de ces formes reste pour le moment non élucidé.

À l'image de STAT1, l'inhibition de Crm1 mène également à l'accumulation nucléaire de STAT2, mais STAT2 est également présente sous forme non-phosphorylée dans le noyau[99,100]. L'import nucléaire de STAT2 a montré comme étant lié à IRF9 (*Interferon regulatory factor 9*)[101]. La localisation de STAT2 et STAT1 est considérablement affectée par la stimulation par l'interféron de type I ou III, qui promeuvent la formation d'un hétérodimère STAT2 : STAT1[90,91] dont l'import est régulé au sein du complexe ASGF3 (*Interferon stimulated gene factor 3*). L'import nucléaire de STAT2 est ainsi régulé avant et après sa phosphorylation par deux mécanismes distincts.

STAT3 est présente sous les deux formes phosphorylée et non-phosphorylée, et son import nucléaire semble indépendant de son statut de phosphorylation[102,103]. Le motif porteur du signal d'import nucléaire a été situé dans le domaine CCD[104]. Ce fragment reconnait les importines α3 et 6[103], qui peut ainsi agir à la fois sur la forme phosphorylée et non-phosphorylée de STAT3. La forme phosphorylée de STAT3 peut également se lier aux importines α5 et 7[105]. STAT3 subit une navette incessante entre les compartiments nucléaires et cytoplasmiques[106,107], dont la sortie du noyau est finement reliée à Crm1[108]. L'inhibition de Crm1 n'abolit cependant

pas totalement la sortie de STAT3 du noyau[103], indiquant que d'autres protéines sont impliquées dans ce processus.

STAT4 suit un schéma différent de STAT3: la forme non-phosphorylée est trouvée dans le cytoplasme uniquement, alors que seule la forme phosphorylée est retrouvée dans le noyau[109,110]. Son import nucléaire requiert la présence d'un fragment du DBD.

Zeng et collaborateurs ont montré que l'export nucléaire de STAT5 non-phosphorylée est régulé par Crm1[111], alors que son accumulation dans le noyau est indépendante de son statut de phosphorylation. Une étude a montré que l'import/export de STAT5 se déroule de manière continue[112]. La délétion d'un fragment du CCD permet d'arrêter l'import de STAT5a, même si le peptide correspondant n'est pas suffisant pour introduire la fonction d'import nucléaire[113], ce qui indique que la régulation de l'import nucléaire se fait *via* un motif inhabituel qui s'étale sur plusieurs zones du CCD et permet la liaison à l'importine α3. Le recyclage de STAT5 se fait vers le noyau avec l'implication de Crm1, dont l'inhibition n'élimine pas complètement l'export de STAT5a[113]. Les autres partenaires de STAT5 permettant son export vers le cytoplasme restent à déterminer.

STAT6 partage de nombreuses similarités avec STAT5 et STAT3. La cinétique de son import est ainsi équivalente, pour les formes phosphorylées et non-phosphorylées[114] et la stimulation de la phosphorylation de STAT6 se traduit par son accumulation dans le noyau. La cinétique de l'import nucléaire n'est pas influencée par la phosphorylation mais l'export nucléaire est ralenti. De manière identique à STAT3 et STAT5, le motif porteur du signal d'internalisation se trouve dans le domaine CCD, même si les résidus impliqués ne sont pas les mêmes[114]. Enfin, STAT6 peut se lier à l'importine α3, comme STAT3 et STAT5.

Une revue plus exhaustive des phénomènes d'import et export nucléaire au sein des protéines STATs a été écrite récemment par N. Reich[115].

Ainsi, si 4 STATs (3-5a-5b-6) partagent beaucoup de points communs, toutes les STATs ont des spécificités (en termes de protéines partenaires, de régions de reconnaissance des importines, …). L'hétérogénéité des mécanismes qui en découle pourrait constituer un enjeu important dans la recherche de spécificité de composés actifs ciblant la famille de protéine STATs et plus particulièrement les interfaces des STATs avec d'autres protéines. Les STATs possèdent cependant d'autres interfaces qui permettent d'adopter différents arrangements au cours de son cycle.

6. Les différents arrangements (monomères, dimères antiparallèles, dimères parallèles, hétérodimères, tétramères)

Les protéines STATs coexistent dans le milieu cellulaire sous plusieurs arrangements – monomériques, dimériques, ou tétramériques. De plus, des homo- et hétéro-dimères peuvent exister pour certaines protéines STATs, dont les propriétés sont différentes. Il s'agit ainsi d'une voie de modulation de l'activité des STATs qui implique plusieurs interfaces et des

réarrangements spatiaux notables des protéines mais peu ou pas de changement des structures des domaines. Les interfaces et réarrangements des différents domaines impliquent principalement les domaines du *Core Fragment* et le domaine N-terminal, alors que le domaine C-terminal, ou domaine de transactivation, n'est pas indispensable [37,116].

Dans le milieu cytoplasmique, le paradigme initial a longtemps voulu que les protéines STATs soient principalement présentes dans leur forme monomérique latente non-phosphorylée, avant d'être recrutées pour activation *via* phosphorylation d'un résidu tyrosine crucial. Cependant, des données contradictoires sont apparues et ont montré que des formes dimériques de STAT non-phosphorylée sont présents *in cellulo*. Mao et collaborateurs ont ainsi résolu une structure cristallographique de STAT1 non-phosphorylé mais dimérique dans lequel deux interfaces sont présentes, une entre les domaines N-terminaux et l'autre entre le CCD et le DBD[117]. Dans cette configuration, les deux domaines SH2 sont rejetés aux deux extrémités de la structure ; les arrangements de ce type sont ainsi dénommés anti-parallèles (*cf.* Figure 3). Un dimère similaire a été résolu pour la forme murine de STAT5a, et présente la même interaction entre les deux *Core Fragments*[118]. Malheureusement, les domaines N-terminaux ne sont pas présents dans la construction de la protéine. Une étude par SAXS des formes cytoplasmiques de STAT5 a également indiqué que les formes monomériques et dimériques sont toutes deux présentes, bien qu'il apparaisse que la forme dimérique est minoritaire[119].La présence d'une forme dimérique avant l'activation a également été confirmée pour STAT1[120], et STAT5[118]. Une structure cristallographique de STAT3 non-phosphorylée a été résolu à partir des données obtenues par rayons X, indiquant un arrangement monomérique de STAT3 sans domaine N-terminal en solution, ce qui a été appuyé par une autre étude[121,122]. L'arrangement adopté par STAT3 en présence du ND reste pour le moment sujet à caution, même si des formes dimériques non-phosphorylées ont été trouvées, ainsi que pour STAT1[123]. Cependant, si l'organisation des dimères est identique, des différences sur certains points sont à noter : l'affinité des domaines ND de STAT3 et STAT1 diffèrent sensiblement[116] et des mutations déstabilisant le dimère anti-parallèle de STAT1 n'ont que peu d'effets sur STAT3 [121,122]. Ainsi, si la présence des formes dimériques non-phosphorylées semble être constante pour toutes les protéines de la famille des STATs, les interfaces sont spécifiques à chaque protéine et restent peu étudiées jusqu'à présent. Enfin, des hétérodimères (STAT1:STAT2 et STAT1:STAT3) ont été retrouvés dans des cellules non-stimulées[124,125]. L'équilibre entre ces différents arrangements (monomère, homodimère et hétérodimère antiparallèle) reste de plus très peu étudié, même si plusieurs études rapportent que STAT3 et STAT5 seraient principalement monomériques dans le cytoplasme[118,119,121].

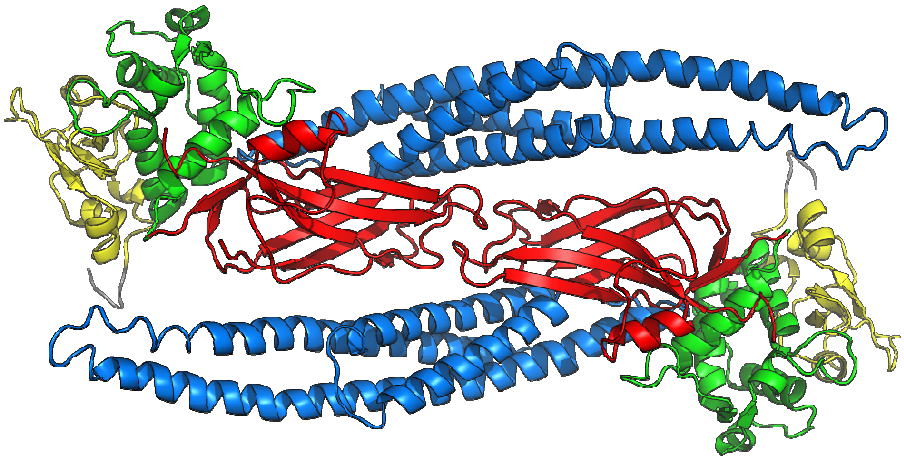


Figure 3: **Représentation en ruban d'un homo-dimère antiparallèle STAT5.** Seul le *Core Fragment* est représenté suivant le schéma de couleur suivant : le CCD en bleu, le DBD en rouge, le LD en vert, le domaine SH2 en jaune.

Si la dimérisation des formes cytoplasmiques non-phosphorylées de STAT5 a récemment été clairement établie, l'existence d'une forme dimérique après activation a été démontrée très tôt suite à la découverte de cette famille de protéines. La dimérisation autour d'un double brin d'ADN des formes actives de STAT existe pour toutes les STATs, et implique l'interaction réciproque du résidu de tyrosine phosphorylée d'un monomère avec le domaine SH2 du second monomère[126]. Cet arrangement où les deux domaines SH2 se font face est appelé dimère parallèle (*cf.* Figure 4), et peut impliquer des monomères différents pour former des hétéro-dimères hybrides STAT1:STAT2, STAT1:STAT3, STAT1:STAT4, STAT1:STAT5, STAT2:STAT3 ou STAT5:STAT6[127–131]. Les structures cristallographiques de ces formes parallèles liées à l'ADN ont révélé une grande similarité de l'organisation des STATs de mammifères[132–134], de même que la structure d'un homologue de STAT chez *Dictyostelium discoideum*[135]. Cette forme dimérique non-liée à de l'ADN montre la même organisation des domaines structuraux de la protéine, ainsi qu'une dimérisation réalisée par l'interaction SH2:phosphotyrosine, mais adopte une conformation étendue qui nécessite un réarrangement structural important afin de pouvoir se lier à l'ADN. Ue tel arrangement n'a cependant jamais été caractérisé à ce jour chez les mammifères.

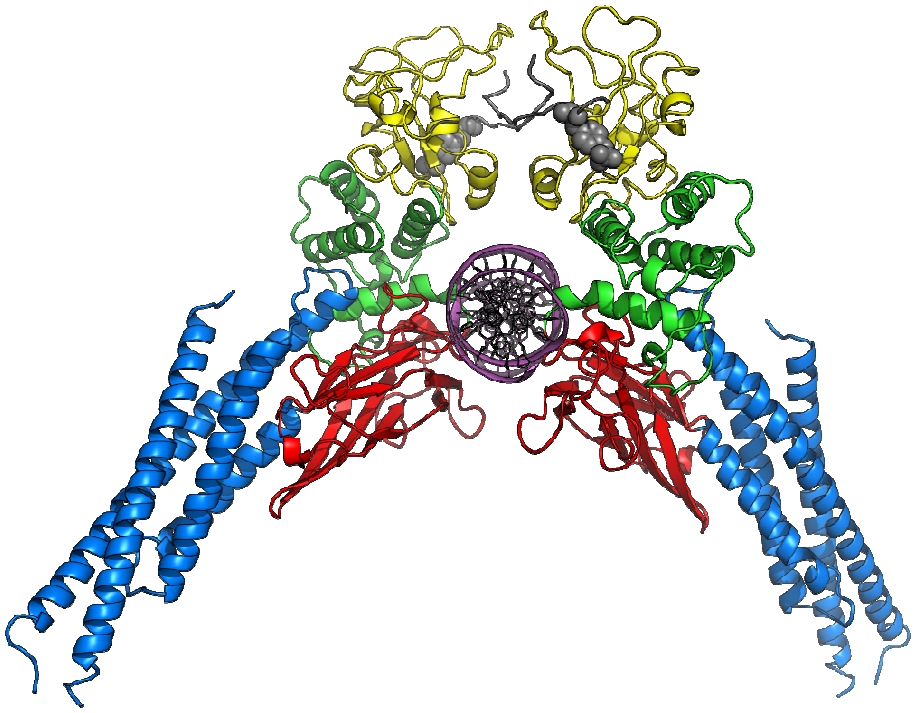


Figure 4: **Représentation en ruban d'un homo-dimère de STAT3 lié à l'ADN :** L'ADN est en violet, et le C*ore Fragment* est coloré comme suit : le CCD en bleu, le DBD en rouge, le LD en vert, le domaine SH2 en jaune et les résidus de phosphotyrosines sont affichées en boules grises.

Au contact de l'ADN, des structures de STATs plus complexes peuvent également se former, *via* la formation de tétramères ou de complexes d'ordre plus élevé. Des dimères parallèles se fixent à l'ADN, sur des sites proches, et forment des contacts par leurs domaines N-terminaux[136–139]. Les facteurs régulant la formation et l'extension de ce type de structure restent peu connus, et les études de telles structures pour la transcription des gènes ont principalement porté sur STAT5. Elles ont révélé que les tétramères peuvent se fixer sur des sites ADN à faible affinité et contrôler l'activité de promoteurs bien définis[140], dont la transcription est fortement réduite lorsqu'on empêche la formation des tétramères par des mutations[141,142]. STAT5 nécessite donc cette capacité d'oligomérisation en présence d'ADN pour remplir l'ensemble de ces fonctions de régulations, physiologiques ou pathologiques dans le cadre de la leucémogenèse[143]

7. Les gènes ciblés par les STATs, les processus biologiques/physiologiques STAT-dépendants et l'implication des STATs dans diverses pathologies

Les protéines STATs sont exprimées dans de nombreux tissus, et sont par conséquent impliquées dans la régulation de très nombreuses fonctions cellulaires. Les principales fonctions régulées par les protéines STATs seront abordées ci-dessous.

STAT1 est exprimée dans de nombreux tissus, notamment dans le cœur, le thymus, la rate et les glandes mammaires où il est retrouvé à de très hauts niveaux[144–146]. Il a été montré que les souris déficientes en STAT1 présentent des anomalies du système immunitaire qui les

rendent très susceptibles à diverses infections. Les souris deviennent insensibles aux interférons α et γ, indiquant un rôle important joué par STAT1 dans la signalisation par interféron[147,148]. STAT2 est exprimé dans la majorité des tissus humains[149]. Les souris déficientes en STAT2 se sont également révélées sensibles aux infections virales en conséquence de la perte de régulation liée à l'interféron de type I[150]. STAT3 est également présent dans la plupart des tissus[151], mais est aussi nécessaire au développement et à la survie des fœtus de souris[152]. STAT3 est impliqué dans la signalisation de nombreuses cytokines, parmi lesquelles IL-6[6,153,154], IL-10[155], IL-21[156,157], IL-27[158] et G-CSF (*Granulocyte colony-stimulating factor*)[159]. La déficience en STAT4 entraîne chez les souris une réponse biaisée à l'IL-12 et l'IL-23, et se traduit par des troubles de la différenciation des lymphocytes Th1 et des cellules *Natural Killer* (*NK*)[3,160]. STAT4 est retrouvée dans les testicules, le thymus et la rate à de forts taux[151,161]. Les deux isoformes de STAT5 peuvent avoir des rôles différents dans certains organes, comme les organes sécrétoires, les muscles ou le cerveau[162–164], et participent aux voies de signalisation impliquant de nombreuses cytokines. Les souris déficientes en STAT5a et b présentent des troubles du développement des glandes mammaires, un retard de croissance, en lien avec le rôle de ces protéines dans la transmission des réponses à la prolactine et à l'hormone de croissance[165–167]. De plus, les souris présentent une anémie fœtale due à une érythropoïèse perturbée et une mort cellulaire augmentée[168]. D'autres implications de STAT5 dans le développement, la différenciation et la survie des cellules hématopoïétiques sont également connues et caractérisées[169–171]. STAT6 est très répandue dans l'organisme [4] et son absence engendre des souris qui sont réfractaires à l'IL-4 et l'IL-13, en association notamment avec des troubles de la polarisation des cellules Th2 et une plus grande sensibilité aux infections parasitaires[172–174].

Les formes non-phosphorylées des protéines STATs peuvent également montrer une activité, distincte de celles brièvement exposées ci-dessus. Ainsi, Cheon et collaborateurs ont montré que les STAT1 et STAT3 non-phosphorylées persistent plusieurs heures après stimulation par les interférons (STAT1) ou l'IL-6 (STAT3)[175] et possèdent une activité de facteur de transcription distincte de celle de leurs formes phosphorylées. D'autres études ont depuis montré l'implication d'autres STATs non-phosphorylées dans la régulation l'hypertrophie cardiaque[176], la régulation basale de gènes activés par l'interféron α[177] ou l'immunomodulation de cellules souches mésenchymateuses[178]. L'activation des STATs (c'est à dire leur phosphorylation) peut également conduire à la modification de la structure de la chromatine : une étude a ainsi montré que l'activation de STAT1 engendre un remodelage du locus du complexe majeur d'histocompatibilité, nécessaire à sa transcription. La forme non-phosphorylée de STAT1 a donc un rôle répresseur vis-à-vis de ce locus[179]. Enfin, de nouvelles études ont démontré l'implication des protéines STATs dans d'autres compartiments cellulaires que le noyau : STAT3 a un rôle mitochondrial et est impliquée dans la respiration cellulaire et la transformation oncogénique dépendante de RAS[180–182], alors que les formes non-phosphorylées de STAT5 sont présentes dans l'appareil de Golgi et le réticulum endoplasmique rugueux[183].

B. Rôles physiologiques et processus biologiques STAT5-dépendants

1. Cellules souches hématopoïétiques

L'hématopoïèse est un processus finement régulé par des mécanismes de rétrocontrôles complexes dont la finalité est de produire des cellules sanguines matures. Les principaux évènements de différenciation cellulaire résultent d'un équilibre subtil entre différents facteurs de transcription, dont STAT5a et STAT5b qui se sont révélés être des régulateurs importants de l'hématopoïèse avec des rôles pléiotropes dans les cellules souches hématopoïétiques[170,184–189], les progéniteurs hématopoïétiques[190–192] et les cellules matures[193–196].

L'analyse du rôle de STAT5 dans différents processus biologiques s'est faite en grande partie par l'observation des phénotypes de souris déficientes pour un gène de STAT5, ou pour les deux gènes[166]. Les souris homozygotes auxquelles il manque les deux gènes STAT5a et STAT5b sont viables, et ont permis de montrer que STAT5 est importante pour la signalisation de l'IL-2, et est requise pour la prolifération des cellules T d'une part, mais également pour produire les cellules *Natural Killer* (NK) de la lignée B[194,197]. Par la suite, le récepteur à l'IL-7 s'est avéré nécessaire à la maturation des lignées B et T[198]. Plus récemment des différences fonctionnelles ont été mises en avant : l'expression d'Erβ (le récepteur à l'estrogène β) est régulée par la prolactine *via* STAT5b (et pas par STAT5a)[199], et des patients déficients pour STAT5b ont montré un nombre réduit de cellules T normales et un déficit de croissance[200–202]. Ces phénotypes confirment l'hypothèse d'un rôle non redondant des protéines STAT5 dans la régulation de gènes importants, comme par exemple Foxp3 dans les cellules T CD4+et CD8+[203–205]. L'activation de STAT5 dans les cellules souches hématopoïétiques est déclenchée par plusieurs autres cytokines telles que l'IL-3[51,206,207], la thrombopoïétine (TPO)[208–212], le facteur de stimulation des colonies de granulocytes (*granulocyte-colony stimulating factor*, *G-CSF*)[213,214] et le facteur de stimulation des colonies de granulocytes et de macrophages (*granulocyte-macrophage colony stimulating factor*, *GM-CSF*)[213,215–217]. La stimulation par ces cytokines, donc l'activation de STAT5, entraîne la production des lignées cellulaires myéloïdes (qui donnent les mastocytes, les granulocytes et les monocytes ainsi que leurs progéniteurs) et lymphoïdes (qui donnent les cellules lymphocytaires et NK) [194,195,218]. La délétion de STAT5 entraîne à l'inverse des défauts dans ces deux lignées. Le rôle primordial de STAT5 dans les cellules des lignées B et T est discuté en détail dans les paragraphes I.B.2 et I.B.3 ci-dessous3, respectivement.

L'érythropoïétine (EPO) est également un stimulateur de STAT5 dans les lignées érythroïdes [168,219,220], où il peut créer une synergie avec le facteur des cellules souches (*Stem cell factor*, *SCF*) afin d'activer STAT5[221]. L'action de STAT5 se situe donc à plusieurs niveaux de l'hématopoïèse, et les éléments du sang périphérique résultent des effets de STAT5 sur les cellules souches hématopoïétiques, les progéniteurs primitifs et les progéniteurs des différentes lignées.

2. *Lignée B*

Le développement normal des lymphocytes B est étroitement lié au récepteur à l'IL-7, comme l'on montré différentes études sur des souris ne présentant pas ce récepteur[171,198,222]. STAT5 a été reliée à ce récepteur lorsque (i) des souris *IL7R*$^{-/-}$ avec une activation constitutive de STAT5 ont montré un retour à la normale du nombre de pro-lymphocytes B et (ii) des souris *STAT5*$^{-/-}$ ont montré un phénotype des cellules B identique à celui des souris *IL7R*$^{-/-}$[171,222]. Il a été montré que *Bcl2* et *BCL-xL* (deux gènes qui favorisent la survie cellulaire) sont directement ciblés par STAT5 dans les lignées de lymphocytes T. Par analogie, deux groupes ont étudié le rôle de *Bcl2* dans les lignées de lymphocytes B et ont montré que ce gène activé de manière constitutive n'entraîne pas de développement de la lignée B chez les souris *IL7R*$^{-/-}$[223–225]. Ces études indiquent donc que STAT5 joue un rôle dans le développement et la différenciation des lymphocytes B, ainsi que dans la survie des progéniteurs des cellules B, en régulant l'expression des gènes requis.

Deux facteurs de transcription ont été identifiés comme produisant des effets similaires à STAT5 et au récepteur à l'IL-7 sur le développement de la lignée B lorsqu'ils sont délétés chez la souris[226,227]. L'expression des gènes *Ebf1* et *Pax5* serait ainsi régulée par la voie de signalisation IL-7R/STAT5, comme l'ont montré de nouvelles études. Si le lien entre l'expression de *Ebf1* et STAT5 a été clairement établi[228,229], la régulation de *Pax5* par STAT5 apparaît moins claire et semble indirecte. Ainsi, si des sites de liaison consensus pour STAT5 sont présents au niveau du gène de PAX5[230,231], une étude n'a pas détecté la liaison de STAT5 dans les régions promotrices de *Pax5*[232], même si une forte expression de STAT5 induit de hauts niveaux de *Pax5*[233].

STAT5 serait également impliquée dans le processus de réarrangement des *loci* des chaînes lourdes et légères de l'immunoglobuline (Ig), un phénomène important au cours du développement et de la maturation des lymphocytes B. Cependant, le rôle exact de STAT5 reste discuté. STAT5 se lierait aux sites promoteurs de la région variable de l'immunoglobuline[234] et son activation constitutive s'accompagne d'une augmentation du réarrangement des chaines lourdes[235]. Des résultats contradictoires présentés par Busslinger et collaborateurs ont remis en question les précédentes assertions[236], et ces auteurs ont conclu que l'unique rôle de STAT5 est de promouvoir la survie cellulaire pendant le développement des progéniteurs des lymphocytes B, et ont impliqué *Mcl1*. Certaines données semblent indiquer que STAT5 n'est présente qu'à de faibles taux dans les progéniteurs des cellules B, et servirait à induire l'expression d'*Ebf1* et de *Pax5* en amont de ce stade de progéniteurs.

D'autres études ont porté sur la régulation du réarrangement des chaînes légères d'Ig et ont montré le rôle important de la voie de signalisation IL-7R/STAT5[237,238], STAT5 étant impliqué dans la répression des réarrangements des chaînes légères d'immunoglobuline au cours des premières phases du développement des lymphocytes B. Cette hypothèse est soutenue par la capacité de STAT5 à se lier directement sur un site promoteur du locus des chaînes légères de l'immunoglobuline, et par le fait que la délétion du gène *Stat5* entraîne un réarrangement des chaines légères de manière prématurée dans les cellules progénitrices de la

lignée B[236]. Ainsi, et même si le mécanisme exact n'est pas identifié, STAT5 a un rôle central dans les réarrangements des gènes des chaînes lourdes et légères de l'Ig, et participe donc activement au processus de maturation, différenciation et survie des cellules de la lignée des lymphocytes B.

3. Lignée T

Comme brièvement abordé dans le paragraphe I.B.11, STAT5 est profondément impliquée dans de multiples aspects du développement et de la fonction des différents lymphocytes T. Tout comme les cellules de la lignée B, la différenciation des cellules T est liée au récepteur de l'IL-7 : lorsqu'il est absent chez des souris *Il7r*$^{-/-}$, on observe une réduction notable des cellules CD8+[198], qui peuvent être rétablies lorsque la souris porte une forme constitutivement active de STAT5[239]. D'autres travaux sont venus étayer ces résultats. Il a ainsi été montré que l'expression de récepteur à l'IL-7 est inversement corrélée à l'expression du couple de récepteurs de surface TCR/CD8[240]. Un modèle a ainsi été proposé, dans lequel les progéniteurs présentant des récepteurs des cellules T (*T cell receptor*, *TCR*) à haute affinité se différencient en thymocytes CD4+, alors que les progéniteurs à TCRs de faible affinité montrent une forte expression du récepteur à l'IL-7, poussant à une expression augmentée de CD8 et à la différenciation en thymocyte CD8+. Un aspect intéressant de ce modèle est qu'il explique l'ajustement de la force du signal de TCR/CD8 et de la signalisation qui en découle, qui favorise l'homéostasie des cellules T sans entraîner de phénomènes d'auto-immunité.

Le mécanisme moléculaire sous-jacent à cet ajustement du signal a été caractérisé lorsque la différenciation vers la lignée CD8 a montré *in vitro* et *in vivo* une dépendance au signal IL-7R/STAT5 suite à l'induction de l'expression de *Runx3*, un des principaux régulateurs de CD8 dont l'expression est normalement induite par les cytokines[241]. De nombreuses données peuvent être extraites de ces résultats :

- des études préparatoires ont montré que l'absence du gène *Stat5* engendre des effets différents de l'absence du récepteur à l'IL-7 ; ce phénomène a par la suite été imputé à un recouvrement des fonctions de STAT5 et STAT6 (STAT6 est en effet activée de manière aberrante par l'IL-7 en l'absence de STAT5),

- le signal dépendant du récepteur à l'IL-7 peut être suppléé par l'introduction d'un transgène *Bcl2*, qui rétablit le développement des thymocytes CD8+. Ce résultat indique que STAT5 pourrait n'avoir d'effet que sur la survie des thymocytes CD8+. Plus important, ces études ont montré que la force du signal TCR/CD8 contrôle l'expression de la chaîne α du récepteur à l'IL-7, et que la signalisation qui découle du couple récepteur à l'IL-7/STAT5 est suffisante pour orienter la différenciation des thymocytes CD8+,

- le nombre normal de thymocytes CD8+ n'est pas établi par l'expression de *Runx3* en l'absence du signal IL-7R/STAT5.

STAT5 joue donc un rôle majeur dans le développement des thymocytes de par ses effets prolifératifs et de survie cellulaire. Enfin, la voie de signalisation IL-7R/STAT5 est nécessaire

dans les cellules T CD8+ matures pour maintenir les taux d'expression de *Runx3* et *CD8*, et intervient dans la réponse des lymphocytes T CD8+ cytotoxiques[242].

Les lymphocytes T CD4+ appelés cellules Th2 sont les principaux acteurs de la réponse immunitaire dirigée contre les infections parasitaires et sont caractérisées par une forte production d'IL-4. STAT5 a été impliquée dans la différenciation de ces cellules suite à une étude qui a montré que l'expression ectopique de STAT5 induit le déplacement de la différenciation des cellules T vers le phénotype Th2 plutôt que Th1, et s'accompagne d'une diminution de la production d'IL-17[243,244]. D'autre part, STAT5 a également été montré comme nécessaire à la production d'IL-2 pour engendrer un effet antagoniste de la différenciation en cellules Th17[245]. Ainsi, l'activation de STAT5 empêche la différenciation vers les cellules Th1 et Th17 et oriente le développement de la lignée cellulaire vers les cellules Th2.

Le développement du système immunitaire a été associé à l'IL-2 suite à l'observation de souris *IL2*$^{-/-}$, qui ont montré l'apparition de défaillances multi-organes liées à un syndrome auto-immun qui engendrent la mort avant 25 semaines de vie[246,247]. L'observation des mêmes résultats pour les souris *IL2ra*$^{-/-}$ et *IL2rb*$^{-/-}$ dépourvues de récepteurs à l'IL-2 a confirmé l'implication de cette cytokine dans le processus de développement des cellules de la lignée T[248,249]. Ces souris meurent cependant plus jeunes que les souris *IL2*$^{-/-}$. Plus récemment, il a été montré que les lymphocytes T régulateurs T_{reg} présentent à leur surface le marqueur CD25 (la chaîne α du récepteur à l'IL-2) ce qui, couplé au fait que les souris *IL2*$^{-/-}$ n'expriment pas ce marqueur, suggère que la mort prématurée des souris déficiente en récepteur à l'IL-2 pourrait être due à l'action de l'IL-2 sur le développement et/ou la survie des cellules T_{reg}. L'implication de STAT5 dans les processus de régulation du développement des cellules T_{reg} a été mis en avant par différents groupes au cours de l'année 2003[239,250,251]. Il résulte de ces études que l'activation de STAT5 est associée à l'augmentation du nombre de cellules T_{reg} et que ces cellules ne peuvent se développer en l'absence de STAT5[203,205]. L'ensemble de ces données indique clairement que STAT5 est nécessaire et suffisant à la signalisation en aval du récepteur de l'IL-2 au cours du développement des lymphocytes T régulateurs. Le rôle de la signalisation IL-2/STAT5 a été davantage étudié et relié à l'expression de Foxp3 au cours du développement des cellules T_{reg}[252,253]. Plus précisément, l'activation (constitutive ou non) de STAT5 induit l'expression du gène *Foxp3* dans la seconde phase de maturation des cellules T_{reg}, et conduit à la production rapide de cellules matures *in vitro*, la conversion des cellules progénitrices vers des cellules matures étant plus rapide dans le cas d'une activation constitutive de STAT5b[253–255]. L'activation de STAT5 n'est par ailleurs pas observée dans la première phase de ce processus, indiquant que l'activation de STAT5 et de Foxp3 n'est que transitoire[256,257]. La relation entre STAT5 et développement des cellules T_{reg} a été récemment exposée dans un article de revue[258]. Enfin, la différenciation entre lymphocytes T régulateurs (T_{reg}) et effecteurs (T_{eff}) a été attribuée à la balance entre les voies de signalisation mTOR et STAT5, cette balance étant régulée de manière réciproque[259].

4. Métabolisme oxydatif

Ces dernières années, plusieurs études ont démontré le rôle que peuvent jouer STAT3 et STAT5 dans la régulation de la production des espèces réactives de l'oxygène (*Reactive oxygene species*, *ROS*), un groupe de molécules regroupant notamment l'ion superoxyde O_2^- et le peroxyde d'hydrogène H_2O_2. En 2009, une première preuve de ce rôle a été apportée lorsqu'une fonction mitochondriale non-canonique de STAT3 a été découverte : la phosphorylation de la tyrosine et la présence du DBD n'étant pas nécessaire[182]. À l'inverse, la tyrosine 694/699 de STAT5a/b doit être phosphorylée pour permettre la translocation vers la mitochondrie[260]. STAT5 a été localisée dans la mitochondrie après stimulation de cellules T par IL-2, et coïncide avec le déplacement du métabolisme vers la glycolyse aérobie observé lors de la stimulation cellulaire par des cytokines. Au niveau moléculaire, HIF-2α est présenté comme un gène cible de STAT5 dans les cellules souches hématopoïétiques, et la diminution de son expression induit une diminution de la multiplication des cellules souches[261]. Enfin, ce changement a également été observé dans les cellules de la lignée T[262].

5. STAT5 dans le développement des glandes mammaires

Plusieurs protéines de la famille STATs sont nécessaires au bon développement des glandes mammaires, chaque STAT étant impliquée successivement, à différentes étapes, soulignant le rôle différent joué par chacune de ces protéines dans la transmission des signaux cellulaires (*cf.* Figure 5).

Le développement des glandes mammaires au cours de l'embryogenèse n'est impacté que par la protéine STAT3, dont la délétion génique conduit à la mort de l'embryon en début de développement[152]. Le mécanisme exact de ce phénomène n'a pas encore été élucidé, mais il apparaît possible que STAT3 joue un rôle majeur dans le maintien et la multiplication cellulaire des cellules souches mammaires, à l'image des cellules souches embryonnaires[263]. À l'inverse, les autres protéines STATs ne sont pas indispensables à la formation des glandes mammaires au cours du développement embryonnaire et chez le jeune individu. Cependant, les STATs 1, 3, 5 et 6 ont une importance cruciale au cours de la grossesse, et sont exprimées pendant la gestation et de la lactation.

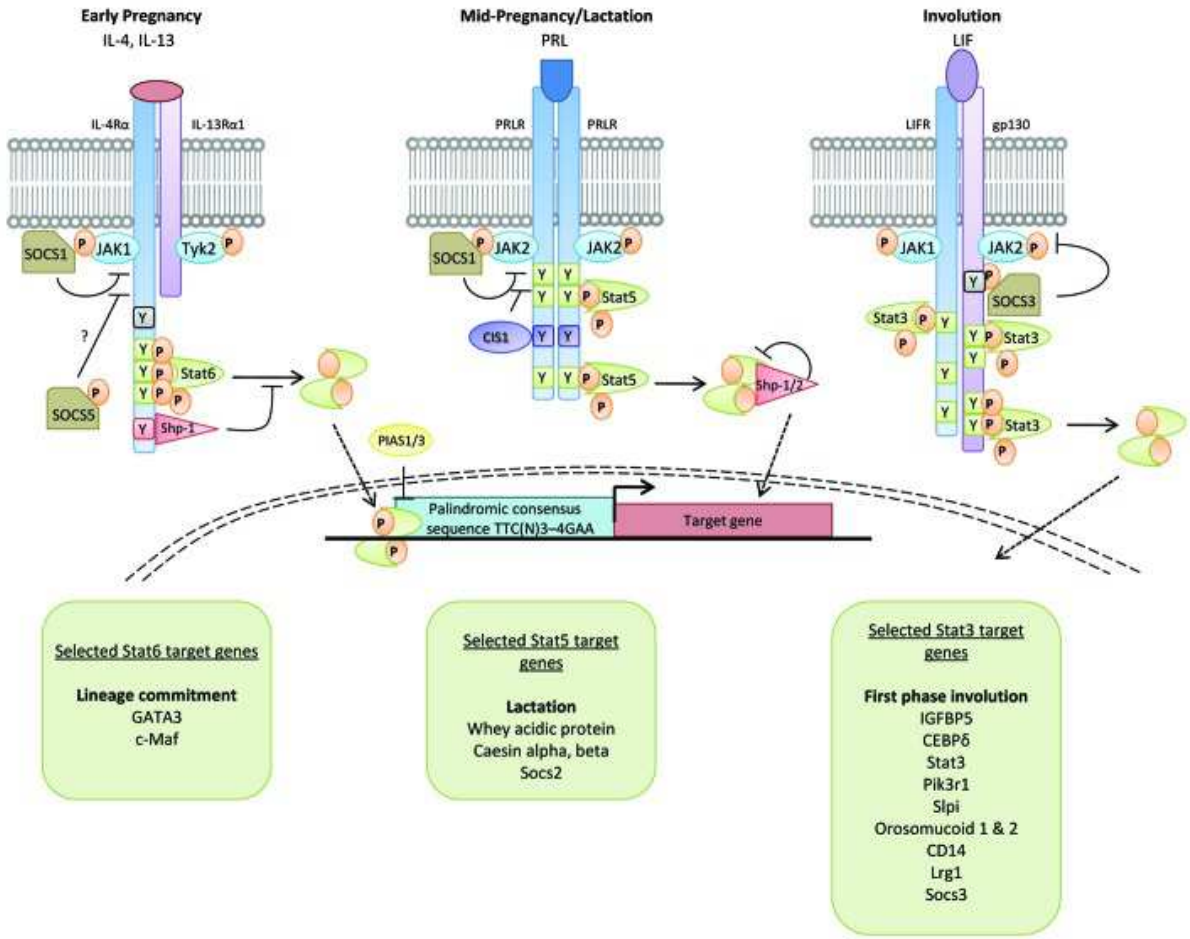


Figure 5: **La voie de signalisation JAK/STAT au cours du développement des glandes mammaires.** Un ensemble de gènes cibles est indiqué pour chaque STAT (STAT6, STAT5 et STAT3) dans les encadrés, en réponse aux stimulations par IL-4/IL-13, prolactine (PRL) ou au facteur d'inhibition de la leucémie (LIF, *leukemia inhibitory factor*), aux différents moments du cycle de développement. Figure reproduite avec la permission des auteurs[264].

Au départ, STAT5 était essentiel pour la régulation de l'expression des gènes codant pour les protéines du lait : en 1991, Watson et collaborateurs montraient que le gène de la β-lactoglobuline (BLG) du mouton présente trois sites de fixation de STAT5 dans sa région promotrice[265]. Des études transgéniques réalisées *in vivo* ont par la suite démontré que l'expression de la protéine BGL nécessite ces motifs[266]. Le lien entre STAT5 et d'autres protéines du lait a ensuite été établi. Le gène de la protéine acide du petit-lait (*Whey acidic protein*, *WAP*) a des sites de liaison de STAT5 dans sa région promotrice, dont la mutation réduit fortement (90%) l'expression du gène de la protéine WAP[267], et l'expression de la β-caséine dans les cellules épithéliales mammaires est également en partie régulée par STAT5[268]. L'activation de STAT5 est principalement liée à la stimulation par la prolactine (*cf.* Figure 5, milieu), et est fortement liée à la régulation de l'expression des gènes des protéines du lait. Des souris KO pour le récepteur à la prolactine ont montré une incapacité à induire l'alvéologenèse au cours de la gestation, ce qui indique que STAT5 est également impliquée dans le contrôle du développement des cellules de l'épithélium alvéolaire[269].

Les effets liés à STAT5a et à STAT5b peuvent être en partie isolés. Une étude a ainsi montré que STAT5a est requis chez la souris pour le développement lobulo-alvéolaire et la lactogenèse[270], tandis que la défection de STAT5b entraîne des défauts de croissance chez la souris[167]. La délétion combinée de ces deux protéines a montré que certaines de leurs fonctions se chevauchent (réponse à l'IL-3 et au GM-CSF, par exemple), mais également qu'elles regroupent la plupart des réponses physiologiques à l'hormone de croissance ou à la prolactine[166]. D'autres études ciblant l'un ou l'autre des gènes STAT5a ou STAT5b à différents moments du développement ont montré que STAT5 est nécessaire à la prolifération, à la différenciation et à la survie des cellules épithéliales[165].

Enfin, Yamaji et collaborateurs ont montré que STAT5a est nécessaire et suffisant au niveau des cellules souches mammaires pour générer les cellules progénitrices luminales[271]. Le mécanisme de la régulation de STAT5a n'est pas encore parfaitement élucidé mais implique *Elf5*, un facteur de transcription dont l'expression est régulée par STAT5a et STAT5b[271], et qui régule l'expression de STAT5a[272]. Ces données semblent donc indiquer qu'une boucle de régulation positive entre *Elf5* et STAT5 existe. Cependant, une étude de 2012 a montré que la présence d'un seul gène de STAT5 est suffisante pour induire la transcription de *Elf5*[273].

Plus récemment, de nouvelles protéines ont été impliquées de manière plus ou moins directe dans la régulation de STAT5 au cours du développement des glandes mammaires. Ainsi, *Miz1*, un facteur de transcription, est impliqué dans la régulation des deux protéines nécessaires à l'activation et l'import nucléaire de STAT5[274]. Enfin, un lien entre la voie JAK/STAT et Akt a été mis en avant par Schmidt et collaborateurs[275].

6. Autres fonctions de STAT5 (régulation du métabolisme des adipocytes, impact de STAT5 dans physiologie hépatique, rôle dans le Système Nerveux Central)

STAT5 est une protéine pléiotrope, et joue un rôle important dans de nombreux processus physiologiques et pathologiques. Si STAT5 est principalement reliée aux fonctions évoquées ci-dessus, elle intervient également de manière moins centrale dans d'autres tissus/types cellulaires car elle reste liée aux voies de signalisation hormonales notamment. Ainsi, plusieurs études ont souligné le rôle la voie de signalisation JAK/STAT qui est impliquée dans la libération hormonale au sein de certaines structures du système nerveux central.

Les souris avec un défaut de STAT5 ou de STAT3 dans ces structures présentent des anomalies neuroendocriniennes qui se traduisent par obésité, diabète et stérilité[276,277]. Ces études ont donc établi un lien entre les protéines STATs et les processus d'homéostasie énergétique et de reproduction. L'hormone principale contrôlant la sensation de satiété est connue sous le nom de leptine. Cette hormone produite dans les adipocytes régule la prise alimentaire et les dépenses énergétiques en se liant à des récepteurs spécifiques de l'hypothalamus. JAK2 est activé par les récepteurs à la leptine et, à son tour, active STAT5, STAT3

et SH2P, ce qui entraîne une augmentation de la production de peptides anorexigènes et une diminution de peptides orexigènes par le noyau arqué de l'hypothalamus[278–280]. De la même manière, dans les neurones synthétisant la gonadotropine (*Gonadotropin-releasing hormone*, *GnRH*), JAK2, STAT3 et STAT5 participent à la transmission du signal afférent[281], donc au processus de production de l'hormone lutéinisante (*Luteinizing hormone*, *LH*) et de l'hormone folliculo-stimulante (*follicule-stimulating hormone*, *FSH*) au niveau de la tige pituitaire. Ces deux hormones sont primordiales dans le développement des fonctions de reproduction, et l'ablation de la fonction de JAK2 dans les neurones synthétisant la gonadotrophine induit une diminution de la sécrétion de GnRH, donc de la fertilité[282]. Enfin, la sécrétion de prolactine est régulée par une boucle au cours de laquelle l'activation de STAT5b est requise[283–286].

Au niveau périphérique, STAT5 présente un effet dans la régulation du développement du tissu adipocytaire[166,287]. Cette propriété a été démontrée grâce à l'étude des effets de la délétion de l'un ou des deux gènes *Stat5* menée par plusieurs laboratoires. Les niveaux de STAT5a et b sur des cellules murines pré-adipocytes en cours de différenciation sont en effet augmentés et corrélés à l'apparition du phénotype cellulaire caractéristique des adipocytes ainsi qu'à des niveaux élevés des facteurs de transcription exprimés au cours de la différenciation des adipocytes (PPARγ, C/EBPs)[288]. De plus, l'expression de STAT5a dans des précurseurs multipotents induit l'engagement de ces cellules dans la lignée adipocytaire[289,290], alors que STAT5b augmente les effets de la différenciation induite par STAT5a. Ces résultats indiquent que les effets de STAT5a et STAT5b sont distincts au cours du développement de la lignée adipocytaire. L'hormone de croissance (*Growth* Hormone, *GH*) est la principale cytokine impliquée dans l'activation des protéines STAT5s par les kinases JAKs dans les pré-adipocytes [291–293]. Elle active STAT5 qui induit l'expression de PPARγ afin de promouvoir la différenciation des adipocytes, comme suggéré par les preuves indiquant que STAT5 peut se lier directement aux sites promoteurs de PPARγ et les activer [291,294,295], PPARγ étant un régulateur majeur requis pour la différenciation des cellules adipeuses[296,297]. De même, la liaison de STAT5a et C/EBPβ (un facteur de transcription qui remodèle la chromatine) durant les phases précoces du développement des cellules de la lignée adipocytaire montre bien l'implication de STAT5a dans la préparation des sites promoteurs pour la fixation des facteurs de transcription[298]. Ainsi, STAT5a, et dans une moindre mesure STAT5b, jouent un rôle majeur dans la régulation du développement des lignées adipocytaires en transmettant des signaux pro-adipogéniques dans la plupart des modèles cellulaires.

Une cytokine activatrice de STAT5 est l'hormone de croissance (GH). Il a été montré que cette voie de signalisation régule la croissance post-natale *via* la régulation de l'expression d'IGF-1[167]. La mutation du gène *Stat5b* chez un patient engendre en effet une insensibilité à la GH consécutive à la déficience en IGF-1, ainsi qu'un défaut de croissance[299]. Deux sites de liaisons de STAT5b sont ainsi présents en amont du gène de l'*Igf-1*, démontrant que l'expression de ce gène peut être induite par STAT5[300]. D'autres sites promoteurs au sein du locus ont été localisés depuis, et ont fait apparaître différents profils d'expression[301,302]. L'hormone de croissance impacte également le cycle cellulaire, comme montré par une étude réalisée sur des fibroblastes

murins et des cellules hépatiques : la déficience en STAT5 augmente la prolifération cellulaire et diminue les niveaux de *Cdkn2b* (*cyclin-dependant kinase inhibitor 2B*) et de *Cdkn1a* (*cyclin-dependant kinase inhibitor 1A*), des inhibiteurs de la progression du cycle cellulaire [303,304]. STAT5 présente également les mêmes effets dans d'autres types cellulaires : l'activation constitutive de STAT5a induit ainsi l'arrêt du cycle cellulaire *via* l'activation de SOCS1 et p53 dans des fibroblastes humains[305,306].

La voie GH-STAT5 a été également mise en avant dans la régulation génique dans le foie, organe dans lequel de nombreuses molécules sont synthétisées et où STAT5b est en large excès par rapport à STAT5a[307]. La régulation des acides biliaires, et d'une manière plus générale des hormones stéroïdes, est ainsi influencée par les protéines STAT5s : la voie GH-STAT5 régule en effet l'expression de gènes de la synthèse, d'excrétion et de transport des acides biliaires[308–311]. Le foie est également le lieu de la synthèse des hormones stéroïdiennes, et STAT5 régule l'expression de plusieurs gènes impliqués dans la régulation de ces molécules, comme par exemple les gènes *Hsd35*[312], *Cyp2b9*[310] et *Cyp7b1*[313]. Par ailleurs, l'impact de la délétion de *Stat5b* a montré des effets différents chez le mâle et la femelle[309,314]. La voie GH-STAT5 est enfin associée au métabolisme des estrogènes[315].

C. Implication de STAT5 dans des pathologies

1. La protéine de fusion BCR/ABL1 : un activateur de STAT5

BCR/ABL1 est une protéine hybride constitutivement active dont l'apparition résulte d'un réarrangement chromosomique entre les chromosomes 9 et 22. La translocation réciproque t(9;22)(q34;q11) génère le chromosome Philadelphia (Ph1), qui possède le gène *Bcr/Abl-1*[316,317], et est retrouvé constamment au cours de la leucémie myéloïde chronique (LMC)[318]. L'apparition du chromosome Ph1 dans une cellule souche myéloïde entraîne l'apparition de la LMC, qui se caractérise par trois phases successives : phase chronique, phase d'accélération et phase blastique[319]. En l'absence de traitement, les patients passent de la phase chronique à la phase accélérée puis à la phase blastique, qui présente de nombreuses similarités avec les leucémies aigües.

Le traitement de cette pathologie repose principalement sur la classe médicamenteuse des inhibiteurs de tyrosine kinases (ITKs), dont le premier représentant est l'imatinib, apparu sur le marché au début des années 2000[320]. L'imatinib se fixe dans le site de liaison à l'ATP de la protéine de fusion BCR/ABL1, inhibant son activité constitutive et permettant une amélioration de la prise en charge des patients CML[321]. Ce composé constitue encore aujourd'hui le traitement de première ligne de la maladie en phase chronique[322]. D'autres ITKs ont fait leur apparition depuis, afin de compléter l'arsenal thérapeutique, tels le dasatinib[323], le nilotinib[324], le bosutinib[325] et le ponatinib[326]. Ces molécules présentent le même mécanisme d'action : ils se fixent dans la poche de liaison à l'ATP des tyrosines kinases, empêchant ainsi la phosphorylation (donc l'activation) de ces protéines. Le ponatinib a semblé particulièrement intéressant, puisque

son efficacité a été montrée contre les mutants de BCR/ABL1 résistants aux autres ITKs, et notamment contre la mutation T315I qui est d'apparition fréquente, mais la mise en évidence d'effets secondaires importants du ponatinib au niveau vasculaire a conduit à la suspension des essais cliniques de cette molécule. Si ces produits ont permis d'améliorer le pronostic des patients atteints de LMC, 20 à 30 % des patients développent une résistance à l'imatinib notamment[327]. Récemment, une étude a montré que le niveau d'activation de STAT5 est corrélé à l'apparition de phénomènes de résistance à l'imatinib[328]. Enfin, l'une des principales difficultés à surmonter est l'éradication des cellules souches leucémiques, nécessaire afin d'assurer la rémission complète des patients, et les inhibiteurs de tyrosine kinases ne sont pas capables d'éliminer toutes ces cellules primitives[329,330].

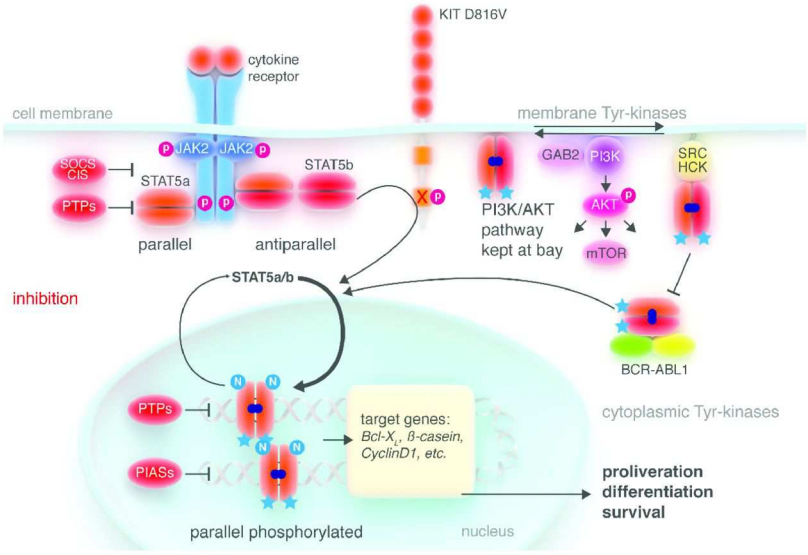


Figure 6: La voie JAK/STAT5 canonique et son détournement par BCR/ABL1 et KIT.

STAT5 est directement phosphorylé sur le résidu de tyrosine Y694 (STAT5a) ou Y699 (STAT5b) par BCR/ABL1 (*cf.* Figure 6), migre ensuite dans le noyau pour (i) réguler le cycle cellulaire, (ii) favoriser l'expression de gènes anti-apoptotiques, (iii) favoriser l'apparition de résistances aux ITKs, et (iv) augmenter le taux d'apparition de mutations *via* l'augmentation des espèces réactives de l'oxygène[331]. Ainsi, il a été montré que STAT5 est requise pour la transformation des cellules et la progression du cycle cellulaire[332], alors que son inhibition induit l'apoptose[333], augmente la sensibilité des cellules K562 à l'imatinib et sensibilise les cellules K562 résistantes à l'imatinib[334]. Des études récentes ont également montré que STAT5 phosphorylée est retrouvée en grande quantité dans le cytoplasme des cellules BCR/ABL1+, et active d'autres voies de signalisation (*cf.* Figure 6)[335,336]. D'autre part, de hauts niveaux de phospho-STAT5 ont été associés à l'apparition de résistances aux ITKs à la fois *in vitro* et *in vivo* et à la progression de la LMC entre ses différentes phases[328]. L'étude d'une cohorte de 50 patients a également montré une corrélation entre les niveaux d'ARNm de *Stat5a* et la fréquence des mutations de *Bcr/Abl-1*, et les niveaux de production des espèces réactives de l'oxygène[337]. STAT5 est donc essentielle pour la progression des néoplasies myéloprolifératives induites par BCR/ABL1 : la perte d'un des deux gènes *Stat5a* ou *Stat5b* entraîne une diminution de la prolifération cellulaire

et un changement phénotypique des cellules vers un phénotype de cellule de leucémie aigüe lymphoblastique à cellules B, alors que l'inhibition des deux *loci* empêche le développement des cellules leucémiques[338]. Récemment, Casetti *et al.* ont utilisé des ARN interférents pour montrer que la délétion combinée de *Stat5a* et *Stat5b* induit l'apoptose des cellules issues de patients atteints de CML, et enlève tout potentiel clonogénique aux progéniteurs des cellules CML[339]. Par ailleurs, les auteurs ont également montré que STAT5a seul est suffisant pour inhiber la croissance de cellules CD34+ de patients atteints de LMC réfractaire à l'imatinib, en augmentant le stress oxydatif et les dommages à l'ADN dans les cellules CD34+ normales ou porteuses du chromosome Philadelphia[339].

Enfin, la voie de signalisation STAT5 interfère par ailleurs avec d'autres voies de signalisation, comme la voie Akt, ce qui pourrait constituer de nouvelles cibles pour le développement de nouveaux traitements. Nous explorons cette possibilité dans une revue récente[12].

2. Les mastocytoses et le récepteur à activité tyrosine kinase KIT

Les cellules mastocytaires dérivent de cellules souches hématopoïétiques de la moelle osseuse qui rejoignent les tissus périphériques *via* le compartiment sanguin. La stimulation du récepteur KIT par le facteur des cellules souches (*Stem cell factor*, *SCF*) induit ensuite la maturation de ces cellules[340]. Le récepteur KIT est un récepteur transmembranaire à activité tyrosine kinase intrinsèque qui dimérise et s'active sous l'influence du SCF[341,342]. L'activation du récepteur KIT induit la phosphorylation de résidus de tyrosine, qui agissent comme des points d'ancrage pour différentes molécules de signalisation et ainsi déclenchent plusieurs voies de signalisation par la catalyse de la phosphorylation de ces molécules[342]. Parmi les voies activées par KIT, on retrouve les voies PI3K/Akt[341], RAS/RAF et JAK/STAT[343–345].

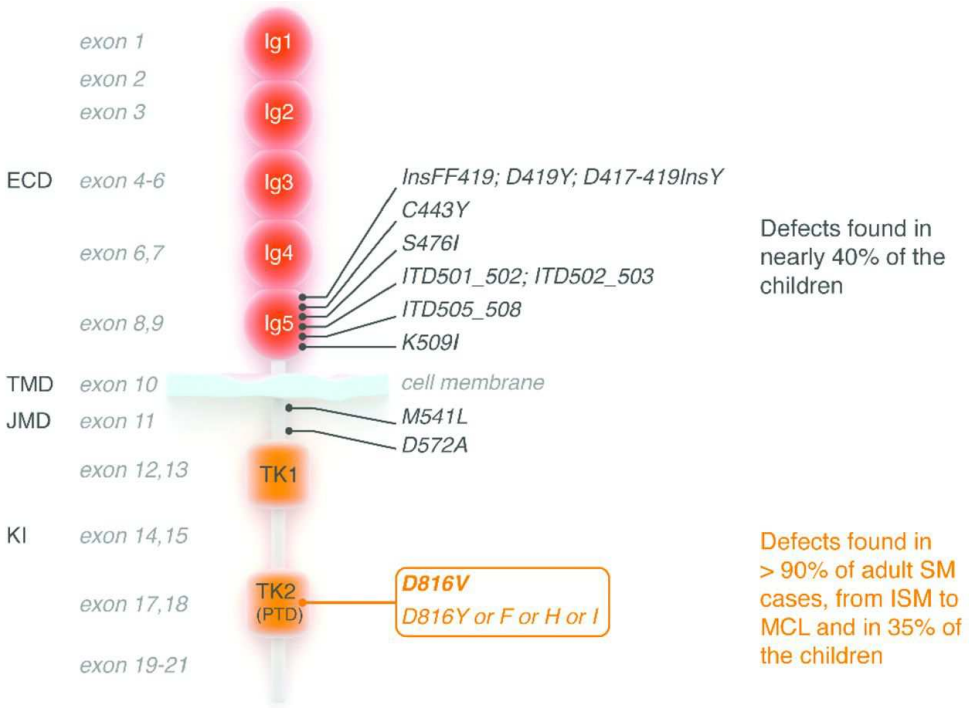


Figure 7: **Représentation de la structure du récepteur KIT, et localisation des mutations oncogènes les plus fréquentes retrouvées au cours des mastocytoses.**

Les mastocytoses sont un groupe de pathologies hétérogènes caractérisées par la prolifération de cellules mastocytaires avec localisation tissulaire multi-viscérale[346], résultant en divers symptômes cliniques. Les mastocytoses peuvent toucher les enfants et les adultes[347], avec des manifestations diverses : les mastocytoses qui se développent durant l'enfance sont généralement limitées à la peau et disparaissent à la puberté, alors que les mastocytoses apparaissant à l'âge adulte sont en générale chroniques, et sont caractérisées par des anomalies systémiques, avec ou sans lésions cutanées[348]. Plusieurs catégories de mastocytoses sont répertoriées par l'Organisation Mondiale de la Santé (OMS), dont les mastocytoses systémiques et cutanées[349]. Les mastocytoses systémiques (MS) sont elle-même subdivisées en mastocytose systémique indolente (MSI), agressive (MSA), leucémie à mastocytes (LM) et mastocytose systémique associée à une maladie hématologique non mastocytaire[347]. Les maladies non mastocytaires associées sont généralement d'origine myéloïde, comme une leucémie aigüe myéloïde (LAM), une LMC ou encore un syndrome myélodysplasique[347].

Dans 90% des cas de MS, une mutation du récepteur KIT est retrouvée (*cf.* Figure 7)[350,351], généralement située sur d'autres sites que les mutations associées aux mastocytoses pédiatriques (*cf.* Figure 7)[352]. Parmi les mutations connues, la mutation *KIT* D816V entraîne l'activation constitutive du récepteur KIT et le recrutement des protéines partenaires associées à des voies de signalisation pro-oncogéniques (*cf.* Figure 8)[344].

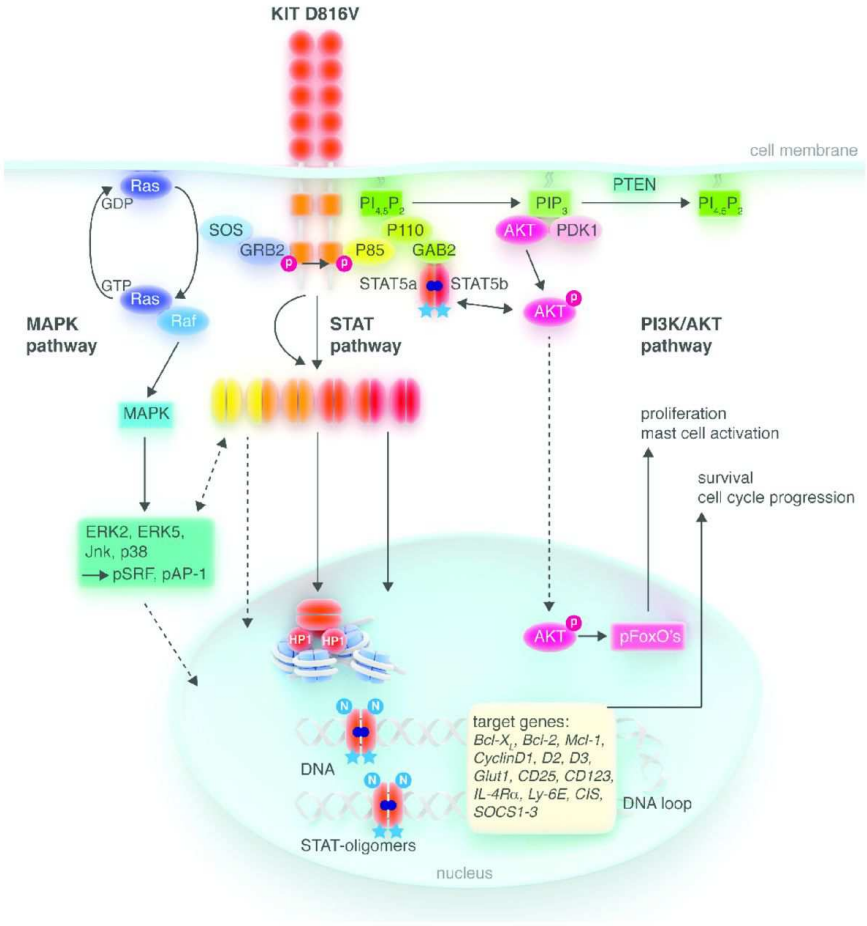


Figure 8: **Voies de signalisation majeures du récepteur KIT D816V muté.** La mutation de KIT en position 816 induit le recrutement des effecteurs des voies de signalisation Ras/Raf/MAPK, JAK/STAT, PI3K/Akt. Ces voies induisent l'expression de gènes de prolifération et de survie cellulaire.

Les traitements appliqués aux mastocytoses sont divers et dépendent essentiellement du type de mastocytose à que présente le patient. Dans le cas des mastocytoses non avancées, le traitement pharmacologique est essentiellement symptomatique, à base d'anti-histaminiques et de glucocorticoïdes[348]. Dans les cas avancés de MS, d'autres composés sont employés pour contrôler l'expansion des cellules mastocytaires. Aucun traitement standard n'a été développé pour combattre les MS agressives, les leucémies à mastocytes ou les MS associées à une maladie hématologique non mastocytaire. Les thérapies à base d'interféron alpha (IFN-α) ont démontré une efficacité variable [353]. La cladribine (2-CdA) peut induire une réponse clinique majeure dans une proportion limitée de patients atteints de MSA[354]. L'usage de ces deux produits est cependant limité par la présence d'effets secondaires tels qu'une cytopénie ou une immunosuppression[353]. Plus récemment, des stratégies basées sur les inhibiteurs de tyrosine kinase sont apparues pour traiter les MSA et les leucémies à mastocytes, du fait de la présence quasi-ubiquitaire de mutations de KIT[355]. L'imatinib est peu indiqué, car la mutation D816V du récepteur KIT y est résistante[356]. Néanmoins, d'autres mutations de KIT restent sensibles à l'imatinib, qui a démontré sa capacité à réduire la charge mastocytaire ainsi que les symptômes des patients atteints de MS[347]. D'autres ITKs (dasatinib et midostaurine) ont montré une activité

inhibitrice vis-à-vis de KIT *in vitro*[355]. Le dasatinib a démontré peu d'efficacité clinique contre les formes agressives de mastocytoses[353]. En revanche, des essais cliniques en cours utilisant la midostaurine semblent montrer une bonne efficacité de cette molécule *in vivo*[357]. Enfin, des stratégies combinant différents composés ont été proposées pour les formes agressives et leucémiques des mastocytoses[358].

D'autres anomalies moléculaires peuvent être retrouvées dans les mastocytoses en dehors des mutations de KIT. Nous avons présenté ces défauts récemment dans un article de revue[351].

3. Rôle de STAT5 dans le développement des leucémies aigües lymphoblastiques

Nous avons vu que le développement et la survie des cellules lymphocytaires sont en partie régulés par STAT5 (*cf.* paragraphes I.B.2 et I.B.3 de ce chapitre). STAT5 est donc logiquement retrouvé à des taux élevés dans les cellules lymphoïdes transformées. Ainsi, des niveaux élevés de STAT5 ont été retrouvés dans plus de 35 à 50% des échantillons cellulaires de leucémies aigües lymphoblastiques (LAL) analysés au cours de deux études [359,360]. Dans ces échantillons, les plus hauts niveaux de pSTAT5 avant traitement ont été retrouvés chez les patients présentant le chromosome Philadelphia, et sont associés à une plus faible survie de ces patients[360]. Ces données suggèrent, sans clairement le démontrer, un rôle majeur de STAT5 dans la transformation tumorale des progéniteurs lymphoïdes B.

Des études menées sur des modèles animaux de LAL BCR/ABL1+ ont apporté cette preuve. En effet, la délétion de STAT5 dans des souris BCR/ABL1+ empêche la transformation des progéniteurs de cellules B alors que l'expression de BCR/ABL1 est détectée et s'accompagne du déclenchement d'une LAL en présence de STAT5[332]. D'autre part, les souris dont l'expression de *Stat5b* est constitutive développent des pathologies semblables à la LAL humaine à une plus haute fréquence que les souris normales[361]. Plus récemment, des études ont montré le rôle important de STAT5 en parallèle d'anomalies des gènes *Ebf1* et *Pax5* pour initier la transformation de ces cellules. La défection de ces deux gènes, qui codent pour des facteurs de transcription importants pour le développement des cellules B, a été retrouvée chez environ 30 % des patients atteints de LAL à cellules B[362]. L'étude de souris présentant à la fois une activation constitutive de STAT5b et une délétion d'un locus de *Pax5* ou de *Ebf1*a montré le développement rapide de LAL fatales. Cependant, aucune différence n'a été notée entre les taux de prolifération des cellules sauvages et des cellules mutées, suggérant que STAT5 n'est pas l'élément initiateur de la transformation. Ainsi, l'initiation de la transformation et la prolifération des cellules transformées résulterait de l'action synergique entre STAT5 et le réseau de signalisation de EBF1 et PAX5, et plus particulièrement la protéine RANKL [363,364].

Récemment, trois différentes mutations de STAT5b ont été trouvées chez des patients atteints de LAL à cellules T : N642H, T648S et I704L. Les effets de ces mutations sont néanmoins différentes, puisque la mutation T648S ne montre aucun effet, alors que les mutations N642H et

I704L induisent une augmentation de la phosphorylation de STAT5 ainsi qu'une élévation de l'activité de transcription de 26 et 17 fois, respectivement[365,366].

4. *STAT5 et les espèces réactives de l'oxygène dans la leucémie aigüe myéloïde*

La physiopathologie des de certaines leucémies aigües myéloïdes (LAM) est liée à la présence de la tyrosine kinase mutée oncogénique FLT3-ITD. Il a été montré que les cellules de ce type de patients présentent un taux élevé d'espèces réactives de l'oxygène (ERO), qui peuvent augmenter le taux d'erreur de réparation de l'ADN et provoquer des cassures double-brin de l'ADN, ce qui induit une instabilité génétique et une mutagénèse augmentée. Ces effets sont associés à des taux élevés de STAT5[337,367], qui ont été reliés à la production de ces espèces réactives de l'oxygène[306]. Dans les cellules de LAM, la phosphorylation de STAT5 permet l'interaction entre STAT5 et Racl, une GTPase, afin de réguler l'activité de NOX (*Nicotinamide adenine dinucleotide phosphoate oxidase*), une enzyme majeure dans la production des espèces réactives de l'oxygène[368]. L'inhibition de FLT3-ITD ou de NOX conduit ainsi logiquement à une diminution des ERO. Par ailleurs, l'inhibition de FLT3-ITD diminue l'activation de STAT5, l'activité de Racl et sa liaison à NOX. De manière intéressante, la diminution de la production des ERO et de l'activation de STAT5 suite à l'inhibition de la sous-unité p22phox de NOX indique l'existence d'une boucle de régulation positive dans laquelle les espèces réactives de l'oxygène produites par la sous-unité p22phox augmentent l'activation de STAT5, qui elle-même accroit la production des ERO *via* la régulation de NOX[368].

5. *STATs et cancer du sein*

Plusieurs études ont fait état de l'activation constitutive de STAT5 dans les tumeurs du sein, notamment dans les tumeurs répondant aux traitements hormonaux[369,370], alors que des souris présentant une forme de STAT5 constitutivement active développent des tumeurs mammaires[371]. Cependant, STAT5 semble être un marqueur favorable de l'absence d'invasion des ganglions lymphatiques[372]. De plus, la prolactine et STAT5 semblent prévenir le passage des cellules tumorales dans le système lymphatique, suggérant un rôle protecteur de la voie de signalisation prolactine/STAT5[373,374].

Un autre membre de la famille de protéines STATs, STAT3, est également impliqué dans le développement du cancer du sein, mais présente des rôles opposés[370]. Ainsi, ces deux protéines sont difficilement dissociables car leur expression est presque toujours concomitante. Néanmoins, les tumeurs présentant ces deux STATs ont un meilleur pronostic que celles présentant l'activation de STAT3 sans l'activation de STAT5. En effet, les tumeurs avec une activation de STAT3 uniquement ont une plus grande probabilité de résistance aux traitements hormonaux, et présentent un caractère plus indifférencié que lorsque STAT5 et STAT3 sont exprimées. La valeur du caractère pronostique de cette caractéristique a été testée dans une cohorte de 295 patients, dont la survie a été examinée et comparée aux profils d'expression

génique[375]. Les patients dont les tumeurs expriment à la fois STAT3 et STAT5 ont ainsi montré une meilleure survie que ceux dont les tumeurs expriment STAT3 uniquement. STAT5 agirait donc comme un modulateur de l'activité de STAT3 dans les cellules tumorales issues d'un cancer du sein. Pour caractériser plus finement les effets de STAT5 et STAT3, Walker *et al.* ont transfecté des cellules de cancer du sein exprimant STAT3 de manière constitutive avec un vecteur vide ou avec une forme de STAT5 constitutivement active [370]. Les cellules possédant les deux STATs ont montré une croissance ralentie, ainsi qu'une sensibilité accrue à différents agents utilisés dans le traitement du cancer du sein et ciblant les microtubules. Les cellules exprimant les deux STATs sont également plus susceptibles de subir une apoptose suite au traitement chimiothérapique[376,377]. La différence des phénotypes des deux populations cellulaires (STAT3 uniquement ou STAT3 et STAT5 co-exprimés) corrèle ainsi avec les différences cliniques observées en termes de survie des patients.

Les hypothèses concernant les mécanismes moléculaires responsables de cette différence sont multiples. Tout d'abord, STAT5 et STAT3 entrent en compétition pour se fixer aux mêmes sites promoteurs, et induisent des effets opposés[370]. L'expression de BCL6 est ainsi régulée négativement par STAT5 et positivement par STAT3, mais en présence des deux STATs, l'effet de STAT5 prédomine et BCL6 est donc réprimé. D'autre part, STAT5 régule l'expression de plusieurs protéines suppresseurs de STATs, dont les SOCS (*cf.* paragraphe I.A.3 de ce chapitre)[378–380]. Ainsi, Walker *et al.* ont montré que les protéines CIS et SOCS3 sont positivement régulées dans les cellules MDA-MB-468. L'activation réduite de STAT3 dans ces cellules dans lesquelles STAT5 est activée ainsi que l'activation des membres de la famille SOCS indique que STAT5 régule négativement l'activation de STAT3.

6. Les protéines STAT5s dans le cancer de la prostate

Au cours des années 2000, des études ont montré que les deux isoformes de STAT5 sont constitutivement actives dans les cellules du cancer de la prostate, au contraire des cellules normales[381,382], et que les *loci* des gènes *Stat5a* et *Stat5b* sont amplifiés au cours de la progression du cancer de la prostate[383]. La prolactine active la voie de signalisation Jak2/STAT5 dans les cellules normales de l'épithélium de la surface. Elle est ainsi impliquée dans l'activation de STAT5 dans les cancers de la prostate[384–386], un facteur majeur de la survie des cellules cancéreuses en culture[381,387,388]. Récemment, la voie prolactine/STAT5 a également été impliquée dans l'amplification de cellules souches anormales dans l'épithélium de la prostate[389]. L'inhibition de STAT5 par suppression du domaine de transactivation (TAD) ou par des oligonucléotides antisens ou des ARN interférents a la capacité d'induire l'apoptose des cellules de cancer de la prostate en culture[381,387,390], et bloque la progression tumorale de xénogreffes sur des souris[387,390]. Il a ainsi été montré que des gènes régulés par STAT5, comme *Bcl-x_L* ou le gène de la cycline D1, le sont négativement dans les cellules de la prostate[387].

STAT5 est également relié à la progression du cancer de la prostate vers des formes plus avancées de la maladie. Il existe ainsi une corrélation positive entre le grade histologique des

cancers de la prostate et la présence des formes phosphorylées de STAT5 dans le noyau cellulaire[382,391]. Les formes nucléaires de STAT5a et STAT5b sont par ailleurs détectées dans la majorité des cancers de la prostate récurrents résistants à la castration. STAT5 agit dans ces cancers de manière synergique avec les récepteurs des androgènes qui potentialisent ainsi l'activité transcriptionnelle de STAT5 par un facteur 10[392]. En accord avec ces données, la délétion de STAT5 entraîne une augmentation de la dégradation du récepteur des androgènes et retarde la progression des cancers de la prostate résistants à la castration *in vivo*[393]. Cependant, STAT5 présente d'autres effets, comme démontré par l'apoptose de cellules ce cancer de la prostate sans récepteur des androgènes[390]. STAT5 régule donc la viabilité des cellules du cancer de la prostate non pas uniquement *via* ce récepteur, mais également par d'autres mécanismes. Enfin, la présence de formes actives (phosphorylées) de STAT5 dans le noyau est un facteur de récurrence précoce pour les patients traités par ablation de la prostate[394].

7. Mutants de STAT5b dans des cas de leucémie à grands lymphocytes granuleux

La leucémie à grands lymphocytes granuleux (*large granular lymphocytes, LGL*) est caractérisée par la prolifération chronique de cellules T cytotoxiques ou de cellules NK (*Natural Killer*), et est souvent associée à d'autres troubles hématologiques tels que des cytopénies sévères[395,396]. Des études ont montré qu'une large proportion (30 à 40%) des patients présente une mutation de STAT3 dans le domaine SH2[397,398] qui augmente la phosphorylation et l'activité transcriptionnelle de STAT3.

Récemment, des mutations de STAT5b ont été isolées chez environ 2% des patients atteints de leucémie à LGL [399,400]. Précisément, les mutations N642H ou Y665F ont été retrouvées. Ces mutations ponctuelles ont été les premières mutations observées chez des patients, et sont situées dans le domaine SH2, à proximité du site de liaison de la phosphotyrosine (N642H) et du domaine TAD (Y665F). Les auteurs de la publication ont posé l'hypothèse d'une stabilisation constitutive de la structure dimérique parallèle liée à l'ADN de STAT5, qui entraînerait l'activation continue des gènes cibles. En accord avec cette hypothèse, les mutants de STAT5b ont montré une activité transcriptionnelle accrue et une phosphorylation du résidu Y699 augmentée. D'un point de vue clinique, le mutant Y665F ne montre pas d'effets significatifs, en accord avec la faible modification de l'activité transcriptionnelle qu'il engendre. D'un autre côté, la mutation N642H est plus active et semble associée à des formes agressives de leucémies à LGL. Afin de conclure de manière plus générale quant à l'agressivité de ces mutations, la recherche d'autres patients qui en seraient porteurs est nécessaire vu le faible nombre de patients inclus au cours de ces études.

8. La leucémie prolymphocytaire à cellules T (LPL-T) et STAT5

La LPL-T est une forme de néoplasie agressive caractérisée par la prolifération de lymphocytes T matures[401,402]. Les patients présentent une progression clinique rapide et sont résistants aux agents chimiothérapiques conventionnels, le pronostic étant généralement mauvais. Des réarrangements chromosomiques sont souvent caractérisés chez ces patients[403,404]. Le séquençage du génome de 50 patients atteints de LPL-T a ainsi conduit à l'identification de plusieurs mutations de STAT5 chez 18 des patients inclus dans la cohorte. Ces mutations sont localisées dans le domaine SH2 : T628S, N642H, R659C, Y665H et Q706L[405].

En conclusion, l'ensemble de ces données cliniques montre que STAT5 constitue une cible privilégiée pour le développement de nouvelles stratégies thérapeutiques. Dans plusieurs maladies prolifératives (LMC, mastocytoses, *etc.*), un inhibiteur de STAT5 permettrait de cibler un élément cellulaire clé dans la transformation et la progression cancéreuse. De nombreuses études montrent ainsi que la délétion des gènes de *Stat5* induit l'apoptose cellulaire des cellules cancéreuses en inhibant la transmission des signaux dépendants de STAT5. De plus, l'efficacité de cette stratégie sur les cellules souches, un problème épineux en oncologie[330], renforce l'idée que STAT5 est une cible pertinente. Cette caractéristique présente définitivement STAT5 comme une cible thérapeutique potentielle, sur laquelle plusieurs groupes de chercheurs ont travaillé.

D. Etat de l'art de l'inhibition de STAT5

STAT5 est une protéine cruciale pour le fonctionnement physiologique de multiples types cellulaires et est impliquée dans de nombreuses fonctions de signalisation et de régulation de la transcription *via* une cascade d'interactions de type protéine – protéine. Cette protéine constitue ainsi une cible potentielle majeure dans la prise en charge de patients dont les pathologies sont associées à ces voies de signalisation dépendantes de STAT5. La plupart des inhibiteurs de STAT5 actuels ne ciblent pas directement cette protéine mais agissent plutôt sur les protéines en amont des voies de signalisation (BCR/ABL1, FLT3 ou JAK) afin de réduire l'activation de STAT5. Néanmoins, plusieurs groupes ont développés des inhibiteurs directs de STAT5, agissant sur la dimérisation de STAT5 ou sur sa capacité à se lier à l'ADN. Enfin, d'autres voies d'inhibition potentielle sont à l'étude.

1. Les inhibiteurs en amont

BCR/ABL1 est l'une des kinases qui entraîne l'activation constitutive de STAT5. L'imatinib est un inhibiteur puissant et sélectif de la kinase Abl qui a montré une efficacité nette, à l'origine d'une amélioration de la prise en charge des patients atteints de leucémie myéloïde chronique[406]. Consécutivement à l'inhibition de BCR/ABL1 par imatinib, le taux de phosphorylation de STAT5 est également diminué. L'apparition de mutations résistantes aux ITKs de première génération, comme l'imatinib, a conduit au développement de nouveaux

composés, dits de seconde génération, comme le dasatinib, le nilotinib ou le ponatinib[407] (*cf.* Figure 9). L'inhibition d'autres kinases induit également une baisse de l'activité de STAT5. Ainsi, l'inhibition par le lestaurtinib, le sorafenib ou le ponatinib (*cf.* Figure 9) de FLT3, un récepteur à tyrosine kinase dont certaines mutations sont associées à l'apparition de leucémies aigües myéloïdes[408,409] induit une diminution de l'activation de STAT5[410–413]. Les protéines JAKs constituent également une cible importante car elles activent directement STAT5. Le ruxolitinib, le CYT387 et le TG101348 (*cf.* Figure 9) sont tous des inhibiteurs ciblant des protéines JAKs qui induisent une diminution de la phosphorylation de STAT5 au cours des syndromes myéloprolifératifs présentant une mutation activatrice de Jak2[414–417].

Compte tenu du mode d'activation de STAT5, le ciblage de ses protéines partenaires situées en amont est une stratégie logique et pertinente. Cependant, cela implique plusieurs limites en conséquence. Les protéines ciblées appartiennent à des familles moléculaires pour lesquelles la spécificité est un problème très présent. Ainsi, les inhibiteurs des protéines JAKs sont actifs sur tous les membres de cette famille, même si leur activité diffère d'une protéine à une autre, ainsi que sur d'autres kinases en dehors de cette famille. Ce manque de spécificité engendre une forte variation des effets d'inhibiteurs chez les patients lorsqu'ils sont utilisés en clinique. D'autre part, STAT5 ne constitue pas la seule protéine partenaire de ces kinases, dont l'inhibition a des effets non souhaités en aval de la protéine ciblée. Enfin, la présence de mutants de STAT5 activés de manière constitutive rend cette stratégie d'inhibition non-productive. Des composés ciblant STAT5 de manière directe constituent donc une approche qui pourrait contourner ces limites.

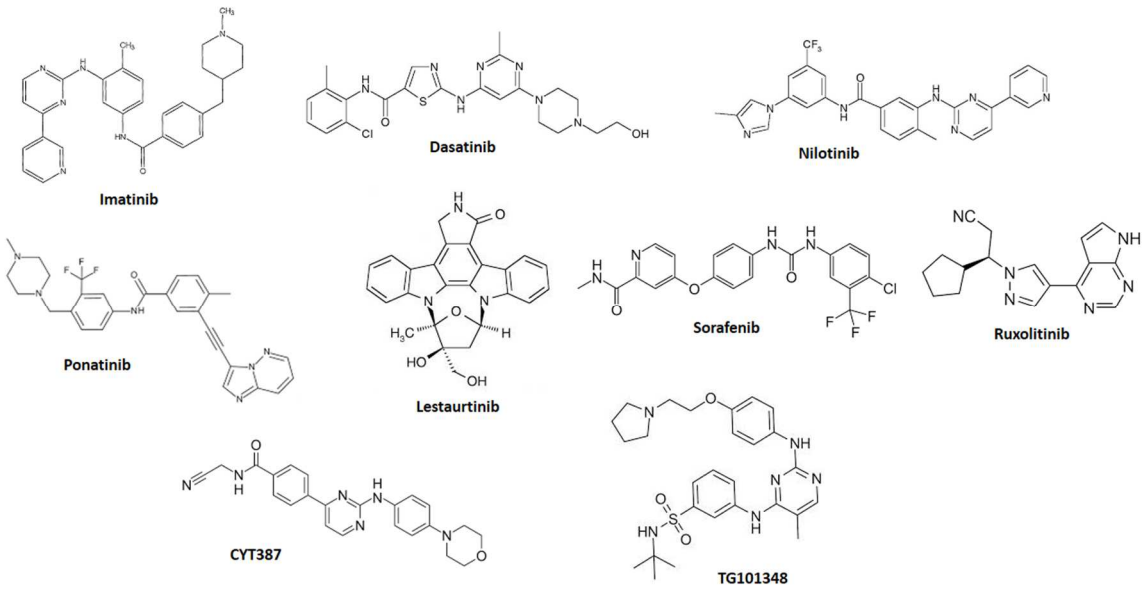


Figure 9: Formules chimiques des inhibiteurs en amont de STAT5.

2. *Les inhibiteurs de la dimérisation de STAT5*

La formation des dimères parallèles est une étape cruciale et nécessaire à la réalisation de la plupart des fonctions de STAT5. Une des façons la plus logique d'inhiber la formation de ce dimère est d'empêcher l'interaction entre le résidu de phosphotyrosine et le domaine SH2. Cette approche a été appliquée avec succès à STAT3[418], puis à STAT5. Une librairie de petites molécules a été criblée à la recherche de composés se liant au domaine SH2 de STAT5, et modulant son activité en conséquence. Une équipe a ainsi criblé une librairie de 17 298 composés et identifié un dérivé chromone de nicotinyl-hydrazone comme étant le produit le plus puissant en termes d'activité inhibitrice. L'activité a été estimée par la mise en évidence du déplacement d'un peptide dérivé du récepteur à l'érythropoïétine du site de liaison du domaine SH2 de STAT5b. Le composé s'est révélé 10 fois plus puissant pour déplacer ce peptide de STAT5b (IC_{50} = 47 µM) par rapport à STAT3 (IC_{50} > 500 µM), et a diminué le taux de phosphorylation de STAT5 dans des cellules de lymphomes après stimulation par IFNα[419]. Une forte concentration (100 à 200 µM) du produit est cependant nécessaire pour observer cet effet.

Plus récemment, une autre équipe de chercheurs a développé plusieurs séries d'inhibiteurs ciblant le domaine SH2 de STAT5. Ils se sont appuyés sur un squelette d'acide salicylique, préalablement détecté comme se liant au domaine SH2 de STAT3[420], pour générer des dérivés inhibant fortement STAT5. Plusieurs molécules, dont le BP-1-108 (*cf.* Figure 10), ont ainsi montré la capacité d'inhiber STAT5 dans des cellules leucémiques K562 et MV-4-11 sans effets cytotoxiques sur les cellules normales. Le BP-1-108 montre une IC_{50} de 17 µM, et diminue l'expression de gènes importants régulés par STAT5 tels que *C-myc*, *Cycline D1* et *D2*, *MCL-1*[421]. Cependant, ces dérivés inhibent également STAT3 et STAT1 à des taux légèrement supérieurs, ce qui en fait des composés peu spécifiques.

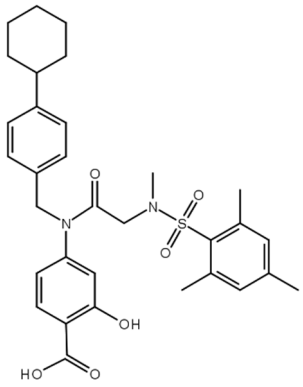


Figure 10: **Structure chimique du BP-1-108.**

Le pimozide (*cf.* Figure 11), qui est un composé utilisé en clinique comme antipsychotique neuroleptique, a montré une activité anti-STAT5[422]. Les cellules leucémiques KU812 ou K562 traitées par le pimozide présentent une diminution importante de la quantité de STAT5 phosphorylée ainsi que de la transcription de plusieurs gènes importants (*Bcl-xL*, *CIS*, *Pim1* ou *Cycline D1*). De plus, le traitement par pimozide induit une diminution de la viabilité et augmente l'apoptose de ces lignées cellulaires de LMC tout en ne montrant aucune toxicité sur

les cellules normales aux mêmes concentrations (10 µM). Enfin, le pimozide a également montré la même activité contre le mutant T315I de BCR/ABL1, un mutant résistant aux inhibiteurs de tyrosine kinase tels que l'imatinib, que sur BCR/ABL1 non-muté.

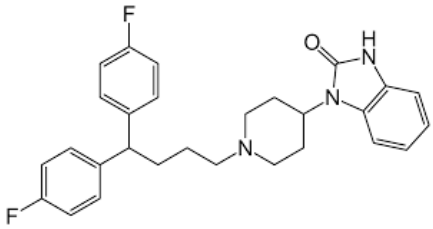


Figure 11: Structure chimique du pimozide.

Un composé issu de la médecine traditionnelle chinoise a également montré une activité anti-leucémique. Des dérivés de l'indirubine, notamment le composé E804 (cf. Figure 12), inhibent STAT5 dans les lignées cellulaires K562 et KCL-22 porteuses de la mutation de BCR/ABL1 T315I résistante à l'imatinib[423]. Cependant, ce dérivé a également des effets sur les kinases de la famille Src (*Src family kinases*, *SFKs*) en inhibant leur autophosphorylation. L'action de E804 se traduit par un blocage de la phosphorylation de STAT5, ce qui induit la régulation négative de l'expression des gènes *Bcl-x*$_L$ et *Mcl-1* dans ces lignées cellulaires. D'autre part, il a été montré que le composé E804 peut induire l'apoptose des cellules K562 et KCL-22 mutées.

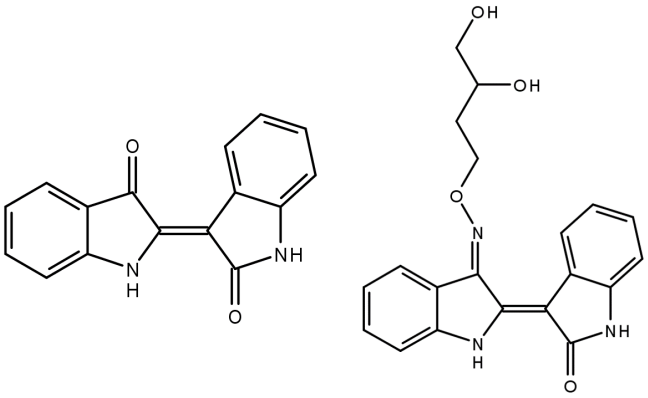


Figure 12 : Structure chimique de l'indirubine (gauche) et du composé E804 (droite).

Plus récemment, un inhibiteur de STAT3 et STAT5 développé par Otsuka pharmaceuticals Co. Ltd. s'est révélé être actif contre une large gamme de cellules hématopoïétiques malignes[424]. Ce composé inhibe fortement la croissance de 20 de ces lignées à des concentrations inférieures à 10 nM, mais également celle de certaines lignées cellulaires issues de tumeurs solides, sans impact sur les cellules normales. Ce composé est entré en phase I de test clinique sur des patients atteints de tumeurs solides, de lymphomes non-Hodgkiniens ou de myélome multiple.

Enfin, l'équipe qui a développé le composé BP-1-108 a synthétisé une nouvelle série de composés dont la sélectivité envers STAT5 est accrue par rapport à STAT3[425]. Le meilleur composé présenté induit le déplacement d'un phospho-peptide (5-FAM-GpYLVLDKW), lié à

STAT5b avec un K_i de 145 nM, alors que le même déplacement avec STAT3 montre un K_i de 144 µM. De plus, le même composé présente une constante de dissociation (K_D) 7 fois plus faible dans le cas de STAT5 (42 ± 4 nM) que chez STAT3 (287 ± 29 nM), sans montrer d'activité contre un jeu de 120 kinases représentatives, dont Hak-1/2, ABL et FLT-3. Enfin, l'activité de ce composé est 3 à 4 fois supérieure à celle des composés de première génération, avec une IC_{50} de 3 à 20 µM sur les lignées cellulaires K562 et MV4;11.

3. Les inhibiteurs de STAT5 ciblant sa liaison à l'ADN

L'activité biologique d'un facteur de transcription est obligatoirement liée à sa capacité à reconnaître et à lier un motif d'ADN. L'empêchement de ce processus constitue donc une stratégie potentielle d'inhibition de STAT5. Des molécules d'oligodésoxynucléotides (ODNs) sont ainsi employées afin de piéger les protéines STAT5, qui vont se lier à ces leurres et se retrouver stockées dans le cytoplasme et incapables d'assurer leur rôle de facteur de transcription[426]. La croissance des cellules K562 a ainsi été stoppée grâce à cette approche[426]. Cependant, l'administration de ce type de molécules reste délicate à maitriser et leur localisation cellulaire restreinte au cytoplasme limite leur pouvoir inhibiteur dans le cas où les protéines STAT5 seraient déjà activées. La vitesse du cycle des deux isoformes de STAT5 est également différente : il faut ainsi 6 heures pour réduire de 90% la quantité de STAT5a activée, alors qu'il faut près de 48 heures pour obtenir le même résultat avec STAT5b[427]. Une séquence aptamère ciblant le domaine de liaison à l'ADN a également été insérée dans un squelette de protéine de la famille des thio-rédoxines et montre une activité anti-STAT5. Plus précisément, l'aptamère interagit avec les protéines STAT5 et modifie leur capacité à être internalisées dans le noyau cellulaire[428]. Cette construction induit ainsi une diminution de la viabilité des cellules tumorales.

Enfin, JQ1 est un composé capable de se lier aux bromodomaines et de ralentir la liaison de ces domaines avec la chromatine et ainsi le fonctionnement des complexes de transcription présents[429]. Les bromodomaines sont des modules génériques présents dans environ 45 protéines pouvant se lier à l'ADN, dont certaines agissent comme coactivateurs de facteurs de transcription. BRD2 est un ainsi un coactivateur de la transcription liée à STAT5 et une étude récente a montré que JQ1 peut inhiber la transcription des gènes dont l'expression est régulée par STAT5[429]. De plus, le composé JQ1 présente une forte synergie avec les inhibiteurs de tyrosine kinase sur l'induction de l'apoptose de cellules leucémiques.

4. Les autres types d'inhibiteurs de STAT5

D'autres stratégies ont été utilisées afin d'inhiber STAT5. Parmi celles-ci, l'emploi d'ARN interférant (ARNi) et d'oligonucléotides (ODNs) anti-sens a montré une certaine efficacité et ces molécules diminuent l'expression de STAT5 au niveau cellulaire[427]. La grande similarité de séquence des deux isoformes de STAT5 permet l'emploi d'un unique fragment d'ARNi ou de nucléotide anti-sens, au prix d'une baisse d'efficacité de la réponse anti-STAT5[164]. Quant aux

constructions à base d'ODNs, elles présentent une meilleure spécificité mais nécessitent l'administration d'un large excès comparativement aux ARNs interférents[164,426].

Au vu de son importance, le domaine N-terminal a également été montré comme site potentiel d'inhibition [430]. Le domaine N-terminal est important pour plusieurs fonctions de STAT5, notamment la formation de complexes tétramériques essentiels par exemple dans la réponse à l'interleukine-2 (IL-2)[142]. Le ciblage de ce domaine pourrait donc constituer dans le futur une stratégie d'inhibition. En effet, il n'existe à ce jour aucun composé ciblant ce site à notre connaissance. Enfin, le cycle de fonctionnement des protéines STATs implique un changement de compartiments, du cytoplasme vers le noyau et inversement. Le blocage de ce transport vers le noyau constituerait une nouvelle stratégie d'inhibition de STAT5, comme suggéré dans une revue récente des stratégies d'inhibition de STAT5 dans les leucémies BCR/ABL1 positive[431].

Depuis sa découverte, STAT5 est impliqué dans un nombre croissant de processus, physiologiques (régulation des cellules souches hématopoïétiques, développement et maturation des cellules des lignées B et T, *etc.*), ou pathologiques (Leucémie Myéloïde Chronique, cancers de la prostate, mastocytoses, *etc.*). Le rôle clé joué par STAT5 dans le développement de ces cancers a ainsi stimulé la recherche de molécules aux propriétés anti-cancéreuses *via* l'inhibition de l'activité de STAT5. Plusieurs stratégies d'inhibition ont ainsi émergées, qui ont prouvé *in vitro* que l'interruption des signaux transmis par STAT5 (progression du cycle cellulaire, inhibition de l'apoptose, *etc.*) empêche efficacement la progression du cancer, y compris de cancers résistants aux traitements actuels. Cependant, ces stratégies nécessitent l'emploi de molécules à des doses incompatibles avec l'administration à des patients. Le développement de nouveaux composés inhibant l'activité de STAT5 et de nouvelles stratégies d'inhibition reste donc un enjeu de taille.

II. Caractérisation expérimentale et théorique des macromolécules biologiques

A. De l'atome à la macromolécule : la hiérarchisation des structures biologiques

L'étude exhaustive des phénomènes biologiques nécessite la connaissance fine des structures des différentes molécules impliquées dans ceux-ci, les principales molécules effectrices étant les protéines. Cependant, les complexes macromoléculaires sont de taille et de composition très variables : la taille des structures (poly)peptidiques varie de quelques angströms (Å) à plusieurs centaines et peuvent impliquer des structures peptidiques, des lipides (membranaires ou non), des molécules organiques, des ions, *etc.* Les entités de nature polypeptidiques peuvent se structurer pour transmettre un signal cellulaire ou alors pour empêcher la transmission de ce signal, afin d'aboutir à la régulation positive (ou activatrice) ou négative (ou inhibitrice) d'un processus physiologique au niveau cellulaire. Les éléments impliqués dans l'initialisation et la modulation des processus cellulaires sont appelés ***effecteurs*** et sont de nature très variables : ligand endogène (cytokine, …) ou exogène (cofacteur, principe actif, …), modifications post-traductionnelles des protéines (phosphorylation) ou mutation génétique (insertion/délétion ou mutation ponctuelle). L'acteur central reste néanmoins la protéine sur laquelle l'effecteur induit une action. La description la plus fine (à l'échelle de taille la plus faible possible) et la plus fidèle possible est donc nécessaire, et les différents niveaux d'organisation des atomes qui forment des structures protéiques constituent la première étape de ce travail.

Le premier niveau d'organisation des protéines est la structure primaire, qui va déterminer la succession des acides aminés composants la protéine – le terme de séquence primaire est également employé (*cf.* Figure 13). De cette séquence primaire, une organisation locale entre résidus apparaît et forme les structures secondaires – hélices, beta-feuillets ou des boucles qui sont des fragments non-structurés. De l'organisation de ces structures secondaires va survenir le repliement tridimensionnel (structure 3D) de la protéine, sa structure tertiaire, auquel la fonction protéique est fortement liée. Enfin, une architecture pluri-protéique peut apparaître afin de former des complexes macromoléculaires (stables ou métastables) : on parle de structure quaternaire. L'ensemble de ces structures est étroitement reliée à la nature des interactions entre les éléments, et à la quantité d'énergie associée à ces interactions.

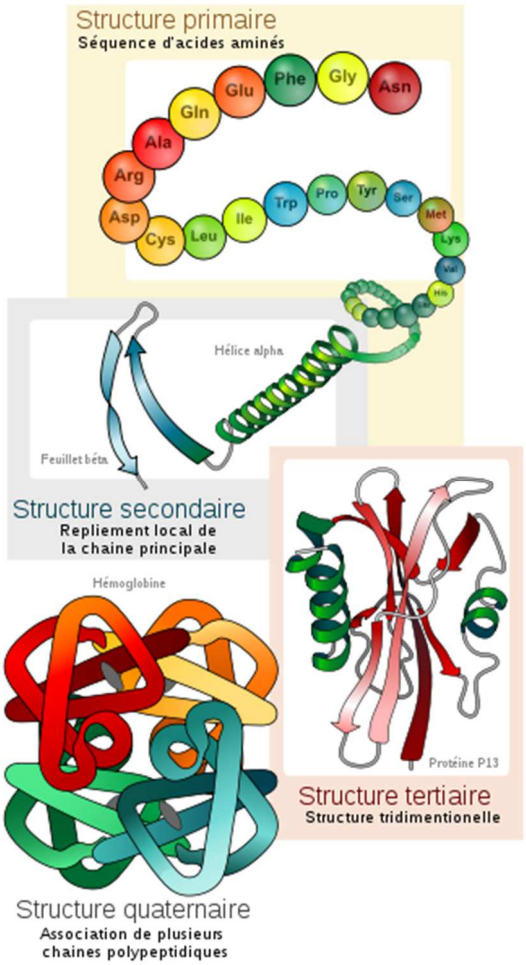


Figure 13: **Les différents niveaux d'organisation des protéines.** (Image issue du site http://www.cours-pharmacie.com/biochimie/structures-des-proteines.html)

La structure primaire des protéines, résultant de la liaison covalente de deux résidus d'acides aminés (liaison peptidique), est associée à une énergie très importante (>100 kcal/mol), est n'est donc formée ou rompue qu'à l'aide d'enzymes hautement spécialisées. L'hydrolyse de telles liaisons étant possible mais extrêmement lente, elles constituent les liaisons les plus stables dans les conditions physiologiques. Les autres niveaux structurels sont liés à des liaisons non-covalentes ayant une énergie plus faible, dont la formation/rupture est en partie possible dans les conditions physiologiques. Ils vont donc apporter une certaine plasticité à la protéine et autoriser les mouvements au sein des complexes : on parle de dynamique conformationnelle ou de transition macromoléculaire.

Les structures secondaires, formées par des liaisons non-covalentes, restent des structures le plus souvent stables car associées à des gains d'énergie importants. Elles résultent de la formation de liaisons hydrogène entre les atomes de la chaîne principale (composée des atomes communs à tous les acides aminés, la chaîne latérale étant formée *a contrario* de groupements spécifiques à chaque acide aminé) de résidus plus ou moins éloignés dans la séquence primaire. Dans les structures de type hélice, l'éloignement des résidus conditionne le pas du tour de l'hélice, et donc son type. Si l'énergie d'une de ces liaisons hydrogène isolée peut

paraître faible comparée aux liaisons covalentes (≈0,5-3,5 kcal/mol[432,433]), la stabilité des structures secondaires est généralement associée à la présence d'un grand nombre de liaisons hydrogène. Cependant, certaines protéines montrent des variations très importantes de leurs structures secondaires, comme la calmoduline dont l'hélice centrale peut se courber.

La structure tertiaire des protéines est associée à différentes forces, covalentes (ponts disulfure) ou non (interactions de van der Waals, …). La formation des ponts disulfure (une liaison covalente entre les atomes de soufre de la chaîne latérale de deux résidus de cystéine) est associée à une énergie de 70 kcal/mol[434] environ. Le gain énergétique, à l'échelle de la protéine ou du peptide, du repliement tridimensionnel associé est cependant plus faible, de l'ordre de 2 à 5 kcal/mol[435,436]. Surtout, contrairement aux liaisons peptidiques, les ponts disulfures sont sensibles notamment aux attaques d'agents nucléophiles et leur rupture a généralement pour conséquence une perte de fonction liée à la perte de la structure secondaire de la protéine. Les forces non-covalentes qui participent à la formation des structures tertiaires et quaternaires peuvent être de deux types : forces de van der Waals ou interactions électrostatiques.

Les forces de van der Waals résultent de l'interaction transitoire entre les nuages électroniques de deux atomes. Généralement de faible énergie (< 1 kcal/mol), elles jouent néanmoins un rôle non-négligeable étant donné le grand nombre d'interactions présentes au sein de complexes de plusieurs milliers d'atomes. Les forces électrostatiques proviennent de l'effet réciproque de deux charges électriques, et sont décrites par la loi de Coulomb. Elles peuvent être attractives (les charges du dipôle sont opposées, ex. : pont salin) ou répulsives (les charges du dipôle sont de même nature), et sont généralement faibles (<2 kcal/mol). Enfin, des interactions apolaires (*i.e.* entre groupements non chargés) surviennent également au sein des protéines, des effets hydrophobes se mettent alors en place, qui peuvent avoir une contribution considérable dans la stabilisation de la structure 3D. La présence omniprésente de molécules d'eau autour de la protéine (l'environnement naturel de biomolécules) peut considérablement pénaliser énergétiquement la protéine si des résidus hydrophobes sont positionnés à la surface. Les résidus hydrophobes vont être tournés vers l'intérieur de la protéine afin de limiter la surface en contact avec l'eau. Ils forment ainsi au sein des protéines un cœur hydrophobe présentant de multiples interactions, faibles individuellement (< 0.7 kcal/mol) mais fortes une fois regroupées (>40 kcal/mol)[437].

La résultante de ces forces explique la structure tridimensionnelle des protéines (forme globulaire ou linéaire, nature des structures secondaires) mais également la formation (ou séparation) de complexes macromoléculaires. La notion de 'complexe macromoléculaire' est omniprésente au niveau cellulaire puisque une cellule est un environnement saturé en protéine, et que chaque protéine est à la fois au centre de **son réseau d'interactions protéique** et à la marge des réseaux de ses protéines partenaires. Réseaux auxquels il faut rajouter les partenaires de natures diverses (hormones, cytokines, sucres, molécules inorganiques, *etc.*). Les interactions entre les différents membres d'un réseau ont ainsi la capacité d'influer l'un sur l'autre, et de modifier leurs propriétés respectives pour moduler leur activité, permettre l'acquisition de

nouvelles propriétés de la protéine (capacité de fixation de nouveaux ligands, …). Ce phénomène est essentiellement dynamique car il s'appuie sur la modification permanente de l'environnement à la fois interne (perturbation du réseau d'interactions entre les résidus) et externe (perturbation du réseau d'interaction entre macromolécule) des différents effecteurs. *In fine*, l'ensemble de ces modifications dynamiques permet la transmission d'un signal cellulaire, conduisant par exemple à l'activation de la transcription d'un groupe de gènes.

L'aspect dynamique revêt une importance primordiale dans la compréhension de la fonction d'une protéine, car sa dynamique reste étroitement associée à sa structure tridimensionnelle. L'intégration de la dynamique moléculaire est l'élément clé pour explorer les relations entre séquence – structures – dynamique – fonctions d'une protéine.

B. La caractérisation expérimentale des structures protéiques

La recherche de la plus petite unité du vivant a longtemps été un centre d'intérêt : la découverte des cellules par Robert Hooke en 1665 a constitué le premier pas qui s'est achevé avec la caractérisation successive des organites intracellulaires (le noyau cellulaire par Robert Brown en 1831, la mitochondrie par Albert von Kölliker en 1857, l'appareil de Golgi par Camillo Golgi en 1883, *etc.*). La biologie structurale a continué à étudier ce pan de la biologie et produit aujourd'hui de nombreuses structures protéiques de taille variée (de petits polypeptides à des ribosomes entiers) et à différentes résolutions (de l'ordre de l'angström Å par rayons X à plusieurs centaines d'angströms par EM). Le développement des bases de données a également joué un rôle primordial dans la publication des structures résolues, et aujourd'hui, plus de 100 000 structures, dont la majorité a été résolue par cristallographie des rayons X, sont accessibles (*cf.* Figure 14).

1. *Détermination des* structures protéiques par cristallographie aux rayons X

Cette technique expérimentale est basé sur l'interprétation de la diffraction de rayons X à travers un cristal obtenu à partir d'une solution protéique, contenant la protéine et d'autres produits, typiquement, un ligand, un inhibiteur ou encore des ions, ainsi que des produits stimulant la solvatation ou l'initiation de cristallisation. La cristallogenèse est le processus qui permet d'obtenir des cristaux et constitue une étape cruciale : plus le cristal sera pur et périodique, moins la carte de diffraction obtenue comprendra de bruit issu des impuretés du signal. Elle représente ainsi une étape souvent limitante dans le processus de cristallographie. Les principaux conditions qui régulent la qualité du cristal sont la pureté de la solution protéique, la concentration, le pH, la température, la présence d'ions, l'emploi d'additifs et savoir-faire de cristallographe… Afin de garder les protéines en solution dans leurs

conformations physiologiques, la cristallogenèse emploie des processus lents tels que la diffusion en phase vapeur pour générer des cristaux dont la taille est de l'ordre du micron (µm) et qui sont composés malgré tout de molécules de solvant aqueux à hauteur de 20-80 %. Des molécules annexes sont ajoutées en règle générale, afin de permettre la formation de contacts cristallins, et de favoriser la stabilité des protéines, améliorant ainsi la qualité de la diffraction. Les molécules au cours de la cristallisation forment une structure périodique, ou un cristal. Les critères fondamentaux sont donc la pureté du cristal, ainsi que l'agencement plus ou moins ordonné des espèces cristallisées (la périodicité) et son taille (taille minimale explorable est 5x10x30 μm^3). Un cristal est une structure solide, composée d'un arrangement ordonné et périodique des éléments qui la composent, dans toutes les directions de l'espace : cet arrangement constitue la maille du cristal, qui est présent de manière répétée par translation dans le cristal. Lorsque la maille est composée de protéines disposées de manière symétrique, une unité asymétrique est constituée du plus petit espace dont la répétition (par rotation, symétrie ou inversion) permet de reconstituer la maille, donc le cristal.

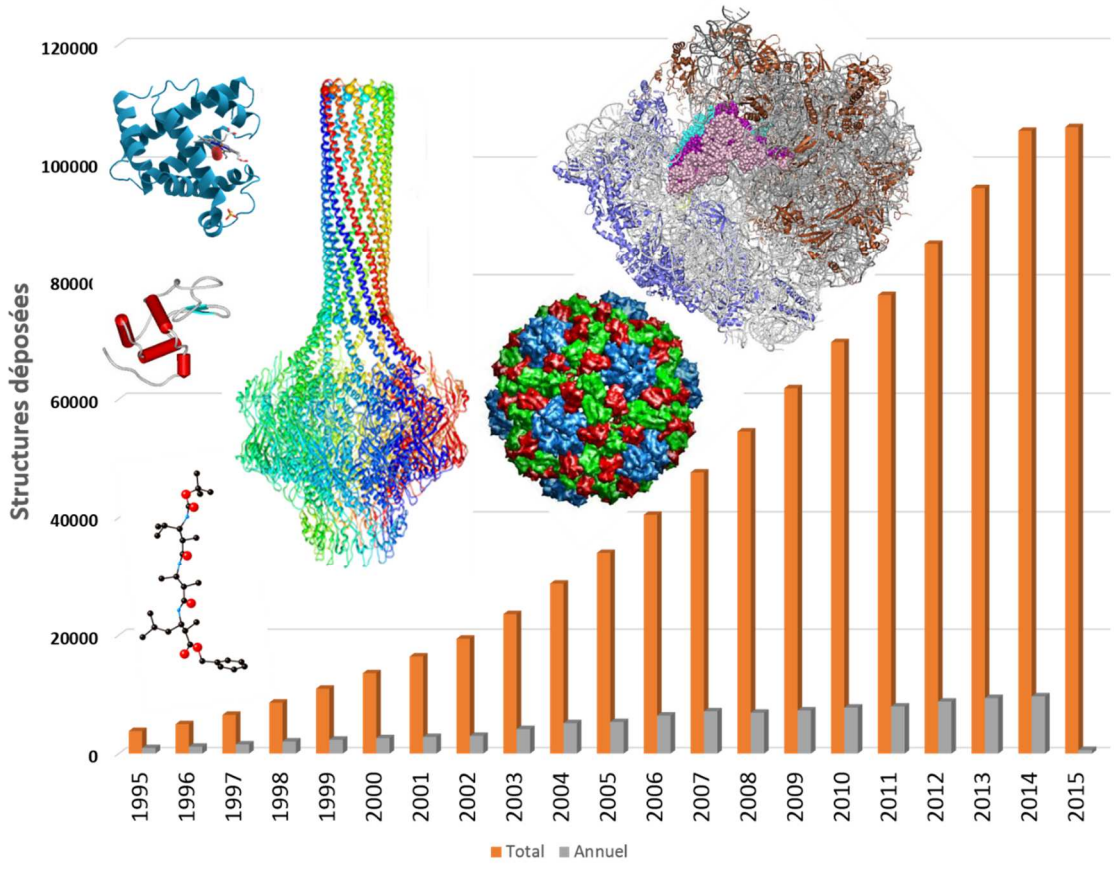


Figure 14 **: Evolution de nombre des structures présentes dans la *Protein Data Bank* et leur complexité croissante.** Le nombre total par année et nombre des structures déposées chaque année montré en orange et en bleu, respectivement. Les structures des simple à complex sont illustré par le peptide, le globine, le lysosime, le bacteriophage, le virus le ribosome.

Les rayons X sont diffractés au contact des nuages électroniques des atomes constituant des motifs répétitifs du cristal. La détection de ces motifs répétitifs par Max von Laue

(récompensé par un prix Noble en 1914) a constitué la base du développement de cette technique qui a abouti en 1957 à la première structure cristallographique[438]. Le faisceau de rayons X qui rencontre le cristal provoquant la dispersion du faisceau lumineux dans des directions spécifiques. La répétition périodique des mailles permet la détection de motif périodique par les détecteurs, qui renseignent ainsi la taille de la maille cristalline et génèrent les cartes de diffraction à partir desquelles est déterminée la carte de densité électronique. À partir de cette densité, la position moyenne des atomes du cristal peut être déterminée, ainsi que leurs liaisons chimiques, leur entropie et d'autres informations. La construction et l'affinement du modèle 'diffractant' sont réalisés en incorporant la composition supposée du cristal (séquence primaire de la protéine) et en comparant intensité de diffraction calculée et observée (facteur R).

La cristallographie permet ainsi d'obtenir les coordonnées atomiques des atomes dont le nuage électronique est suffisant pour diffracter les rayons X, ce qui exclut de fait les atomes d'hydrogène, à l'exception des structures à très hautes résolution (<1 Å)[439]. La précision et l'exactitude des coordonnées, groupées sous le terme de résolution globale, sont exprimées en angströms : les atomes séparés par une distance similaire ou inférieure ne seront pas discernables au niveau de la carte de densité électronique, mais la position des atomes sera généralement déduite à partir de la forme de la carte de densité et des connaissances de la séquence primaire (nature du ou des acides aminés impliqués). Cependant, la résolution globale de la protéine n'est pas uniforme, et elle va grandement dépendre de la régularité des conformations de la protéine du cristal. Les régions les plus flexibles ne peuvent être résolues, car certains fragments des protéines dans le cristal affichent des variations de position trop importantes, produisant une carte des densités électroniques indéchiffrable. Ces régions non-résolues peuvent consister en un simple groupement de chaîne latérale jusqu'à un domaine entier de la protéine. La taille des complexes macromoléculaires (composés de protéines, ligands et/ou acides nucléiques) étudiés varient ainsi de quelques acides aminés à plusieurs milliers, pour des résolutions inférieures à 3 Å[440].

2. La Protein Data Bank (PDB) : une formidable source de structures

La PDB (*Protein Data Bank*, www.rcsb.org)[441] est la base de donnée mondiale qui regroupe les informations concernant les structures tridimensionnelles de molécules biologiques, quel que soit leur nature (protéique ou acide nucléique), leur origine (humaine, bactérienne, …) ou la méthode expérimentale utilisée (cristallographie aux rayons X et au neutron, résonance magnétique nucléaire, cryo-mycroscopie …). D'autres banques de données coexistent (BMRB : *Biological Magnetic Resonance Data Bank*, spécialisée dans le recueil de structures résolues par résonance magnétique nucléaire, *etc.*) mais la PDB constitue actuellement la banque de données de référence de par son exhaustivité (plus de 106 000 structures disponibles début 2015). Cette base de données a standardisé le format des fichiers décrivant les structures, et permet l'intégration de toutes les informations nécessaires à la description de la structure (type d'atomes, leurs coordonnées, facteurs thermiques) et à son

obtention (organisme de la protéine et système d'expression, séquence peptidique de la protéine *versus* séquence peptidique résolue, présence de ligands, …). Enfin, malgré son « exhaustivité », de nombreux systèmes macromoléculaires sont présentés plusieurs fois avec des changements plus ou moins importants (résolution améliorée, présence de ligands différents, …) quand de nombreuses structures et arrangements structuraux restent non-résolus. Ainsi, si la PDB constitue un formidable point d'appui pour toutes les études de modélisation, la limite principale de celle-ci est qu'elle n'offre qu'un regard limité quant à la 'structure exacte' d'une protéine d'intérêt (la grande partie des protéines cristallisées représente des objets modifiés - séquence partielle et/ou bien modifiée) et à la dynamique des protéines, élément essentiel qui gouverne la fonction des protéines et des systèmes macromoléculaires.

3. Analyse, représentation et visualisation des données structurales

En parallèle de la résolution d'un nombre accru de structures protéiques, des outils de visualisation ont été développés afin de permettre la représentation, l'inspection et l'analyse des objets biologiques d'intérêt. Les nombreux types de représentations graphiques permettent de visualiser les structures à différents niveaux, de le superposer pour permettre d'intégrer des analyses systématiques de manière simultanée. Les nombreux logiciels de visualisation (PyMOL[442], Maestro[443], MView[444], Chimera[445], VMD[446], …) jouent donc aujourd'hui un rôle important dans la biologie structurale. Ils permettent une première approche de la structure 3D d'une protéine, des structures secondaires et du positionnement des domaines constituant une protéine. Couplé à l'affichage des séquences et la mise en évidence des résidus clés (selon la littérature et les données cliniques ou biologiques), cette étape de familiarisation avec la protéine constitue généralement la première étape des projets de modélisations. Afin d'étudier ou de comparer le mode d'action de protéines apparentées ou le mode d'activation/d'inhibition de ligands, les logiciels graphiques modernes donne la possibilité de comparer de multiples structures de protéines similaires ou les interactions des ligands liés à une même cible. Au contraire des fichiers de structures protéiques contenant généralement une information limitée (la séquence et les coordonnées atomiques), les outils de visualisation peuvent générer et afficher des informations supplémentaires découlant de la structure (potentiels électrostatiques, charges, …). Un exemple typique est la visualisation des surfaces moléculaires, qui peut être calculées quasi-instantanément avec les moyens informatiques actuels. La représentation des propriétés physico-chimiques sur la surface des molécules, telles que les partenaires le perçoivent *in vivo*, caractérise les systèmes biologiques et sont porteurs d'une information particulièrement pertinente. La superposition d'une même protéine liée à différents partenaires, ou de plusieurs protéines apparentées, permet ainsi d'apprécier rapidement les variations des sites de liaison lorsque leurs surfaces sont affichées.

La dynamique des complexes biologiques reste complexe à se représenter sans les visualiser. La visualisation des mouvements, générés à la volée grâce à des algorithmes dédiés (notamment *via* l'analyse des modes normaux) ou en fournissant un fichier comportant les données de dynamique moléculaire, est un élément important car, en plus de la représentation,

il peut parfois orienter l'analyse de la dynamique et de ses caractéristiques, notamment dans le cadre de projets très exploratoires pour lesquels peu de données structurales initiales sont disponibles. Le développement de ces projets lié principalement à la partie logicielle de visualisation, vont de pair avec une évolution du matériel informatique. L'apparition de nouvelles surfaces de visualisation virtuelle (CAVEs : *cave automatical virtual environments*, écrans 3D…) autorise aujourd'hui la représentation en trois dimensions de structures (et non plus la projection d'un élément tridimensionnel sur une surface) et une immersion accrue, qui peut formuler des concepts ou hypothèses grâce à la perception de détails subtiles difficilement détectables ou non-détectables par ailleurs[447,448]. Ces environnements restent cependant très peu diffusés à travers le monde, du fait de leur coût, si bien que leur utilisation ne concerne qu'une très large minorité des chercheurs. Des modèles structuraux physiques ont également fait leur apparition : reliés à une interface adéquate, ils permettent la manipulation des objets virtuels de manière intuitive et facilitent la communication machine-chercheur ou chercheur-chercheur[449]. Enfin, l'interactivité évolue également d'une façon progressive et la manipulation d'objets avec un retour de forces est maintenant possible, permettant d'explorer en temps réel différentes hypothèses.

C. Modélisation de la structure tridimensionnelle des protéines

Le repliement des protéines dans l'espace a été l'objet de nombreuses hypothèses et études. Christian Anfinsen, pas ses travaux sur la ribonucléase, a montré que la dénaturation de la conformation active, notamment *via* la rupture de quatre ponts disulfure, engendrait la perte d'activité enzymatique[450,451]. Le lien entre la séquence primaire et le repliement spatial des protéines a ainsi été démontré. Quatre-vingt-un pourcents des chaines latérales des résidus non-polaires ont ainsi tendance à se tourner vers l'intérieur de la protéine pour former un cœur hydrophobe (*cf.* paragraphe II.A de ce chapitre), tandis que cette orientation est préféré par 63% et 54% des chaînes latérales polaires et chargées, respectivement[452]. D'autres facteurs ont néanmoins contribuent dans le processus de repliement des protéines : l'action du solvant et des protéines chaperonnes, la présence de cofacteurs, le pH… C'est donc l'environnement cellulaire particulier (protéines associées au ribosome, densité en protéines, *etc.*) qui est responsable du repliement à l'issue de la biosynthèse de la protéine[453], ou de l'absence de structures stables comme dans le cas des protéines ou régions de protéines intrinsèquement désordonnées[454].

Le développement des moyens bioinformatiques et l'accroissement du nombre de structures disponibles (stockées dans les bases de données structurales) ont permis la mise au point de différentes méthodes de prédiction de la structure tridimensionnelle. Plusieurs approches de prédiction coexistent, s'appuyant plus ou moins (voir pas du tout) sur les données structurales disponibles.

La modélisation par homologie (ou modélisation comparative) s'appuie sur les données structurales d'une protéine dont la séquence en acides aminés est proche de la

protéine que l'on cherche à modéliser. Cette méthode repose sur l'hypothèse (validée par la majorité des observations) proposant que deux protéines ayant une bonne similitude de séquence partagent avec une grande probabilité une bonne similarité de structure tridimensionnelle. La différence de structure, représentée comme la déviation de la moyenne quadratique des distances interatomiques des atomes de la chaîne principale, a ainsi été corrélée à la fraction d'acides aminés mutés entre deux protéines homologues[455]. L'utilisation de ce type de méthodes est liée à l'augmentation des données structurales disponibles (au sein de la PDB notamment, *cf.* Figure 14), qui ont permis la modélisation par homologie d'un nombre croissant de protéines, et l'amélioration de la précision des modèles générés. Ainsi, Zhang et Skolnick ont montré en 2003 que des modèles comparables aux structures expérimentales à basse résolution peuvent être générés à partir des données de la PDB (comprenant plus de 23000 structures à cette date), pour des modèles de taille moyenne (*i.e.* moins de 200 résidus)[456].

L'étape initiatrice de la modélisation par homologie est la recherche d'un alignement de séquences homologues. La séquence (ou la fraction de séquence connue) de la protéine à modéliser (la *cible*) est confrontée aux bases de données structurales à la recherche de séquences proches (les *supports*) de la cible, en filtrant les résultats positifs en fonction de la disponibilité des structures associées et de la similarité des séquences. Cette étape, généralement effectuée par un algorithme tel que BLAST[457] ou FASTA[458,459], permet de sélectionner un sous-ensemble de séquences protéiques similaires à celle de la protéine cible et recouvrant tout ou partie de la séquence cible. Différents paramètres (pénalité pour un vide dans la séquence : *gap*, type de matrice de distance utilisée, *etc.*) influent sur la qualité de la sélection, et la sélection du ou des supports finaux se base sur les notions d'identité et de similarité des séquences, du recouvrement des séquences, *etc.* parfois évalués grâce à des scores associés (z-score, E-value). Les informations associées aux structures des supports (résolution de la structure, comparaison des structures secondaires observées pour la protéine support et prédites pour la protéine cible) sont également prises en compte. L'identité des séquences cible-support constitue cependant le critère essentiel : une identité de séquence inférieure à 30% produira des résultats discutables dans le meilleur des cas, alors qu'une identité de séquence supérieure à 50% génèrera des modèles de bonne qualité en règle générale[460]. Pour les séquences à très faible identité, la modélisation par homologie semble être moins adaptée que d'autres méthodes. Alors que plus l'identité de séquence augmente, moins le modèle généré par homologie contiendra d'erreurs. Ces erreurs se localiseront de plus en plus dans les parties très variables de la protéine, comme les chaînes latérales. La précision des modèles générés est également en partie conditionnée par la qualité de l'alignement entre séquence cible et séquence(s) support(s). L'alignement généré par logiciel est choisi, jugé et corrigé par le chercheur et dépend par conséquent de sa capacité à détecter l'alignement optimal à partir des données à sa disposition (structures secondaires prédites de la protéine cible, structures secondaires de la séquence support, conservation des résidus clés, *etc.*).

Une fois l'alignement obtenu, des modèles vont être générés à partir cette information, générant un jeu de coordonnées pour chaque atome lourd (non-hydrogène). Plusieurs

méthodologies peuvent être utilisées : l'élaboration de contraintes spatiales à partir de l'alignement, l'assemblage de fragments conservés, ou encore la recherche de courts segments correspondants à la séquence cible. La première méthodologie – élaboration de contraintes spatiales à partir de l'alignement – est la plus couramment utilisée dans la modélisation par homologie, et consiste à générer un jeu de critères géométriques sous forme de contraintes appliquées aux coordonnées internes de la protéine (distances inter-atomes, angles dièdres). Les contraintes sont ensuite optimisées de manière itérative, ce qui permet également de modéliser les régions désordonnées des protéines comme les boucles[461].

L'évaluation des modèles générés est réalisée par l'estimation d'une énergie, issue de potentiels statistiques ou du calcul des interactions physiques au sein des modèles : plus l'énergie du système est faible, meilleur est le modèle. Les potentiels statistiques s'appuient sur la fréquence d'occurrence des interactions intramoléculaires dérivées de la PDB (ou éventuellement d'une autre base de donnée), et peuvent ainsi produire des scores détaillés (parfois résidu par résidu) en plus d'un score global. Ces potentiels statistiques présentent le défaut d'être moins fiables pour les types de protéines peu représentées dans la base de données initiale. Un exemple typique est l'évaluation des structures de protéines membranaires qui n'étaient que peu nombreuses il y a quelques années. L'élaboration d'un potentiel énergétique se base (1) sur les théories de la mécanique classique où les atomes sont considérés comme des sphères interagissant les uns sur les autres, et (2) sur l'hypothèse que la conformation native des protéines est celle présentant l'énergie la plus faible.

En dehors de la modélisation comparative, **la modélisation par reconnaissance des repliements** est employée[462,463]. Cette méthode ne nécessite pas la connaissance de la structure d'une protéine support homologue, mais s'appuie sur la structure des protéines partageant des structures secondaires et certaines propriétés physico-chimiques (exposition au solvant, hydrophobicité, *etc.*) similaires à celles de la protéine cible. La modélisation par reconnaissance de repliement se base sur le fait qu'il existe un nombre limité de repliements dans la nature[464,465], et que ceux de la plupart des protéines sont déjà représentés dans la PDB. Aussi une protéine ne possédant pas d'homologues de structure connue a de bonnes chances d'adopter un repliement déjà présenté dans les bases de données structurales. Un modèle de la protéine cible est généré (en commençant par la chaîne principale puis la chaîne latérale des résidus) en adoptant le repliement le plus plausible étant donné sa séquence, par enfilement successif des résidus.

Enfin, certaines protéines n'ont pas encore de structures résolues suffisamment homologues pour permettre leur modélisation en utilisant l'une des méthodologies décrite précédemment. **La modélisation *de novo***, modélisation *ab initio* ou modélisation libre, de structures à partir uniquement de la séquence constitue alors une alternative, en se basant sur la propriété que la structure d'une protéine est encodée dans sa séquence primaire[451,466]. Différentes approches sont utilisées : certaines reposent sur la modélisation de fragments de petites tailles (5 à 15 résidus) qui sont ensuite étendus et/ou assemblés[467], tandis que d'autres utilisent des techniques de simulation qui explorent l'espace possible formé par les

conformations de la protéine cible[468], ou encore qui décrivent le processus entier de repliement[469]. Ces dernières simulations restent cependant excessivement rares pour le moment et sont réservées à des protéines de taille moyenne, étant donné le temps de calcul nécessaire pour simuler un processus qui se déroule *in vivo* généralement sur plusieurs millisecondes.

D. L'étude de la dynamique des systèmes biologiques

1. Principes généraux

Tous les processus biologiques reposent sur les propriétés dynamiques des composants cellulaires et les changements structuraux/conformationnels liés à l'exercice d'une fonction biologique donnée prennent place dans des temps variables (*cf.* Figure 15). L'étude et la compréhension de ces processus nécessitent donc de prendre en compte cette dimension temporelle, sans se limiter à l'analyse tridimensionnelle. L'augmentation des moyens de calcul (supercalculateurs de très haut performance) a ainsi permis à l'étude de la dynamique moléculaire par diverses méthodes de se développer au cours des années 70-80 avant de constituer un champ d'étude bien spécifique, ne se limitant pas à la biologie : la science des matériaux a également profité de ces nouvelles méthodes computationnelles. La dynamique moléculaire permet l'étude de processus prenant place dans des échelles de temps allant de la femtoseconde (fs) à la milliseconde (ms), pour des systèmes biologiques de petite taille (peptide[470]) ou très complexes (complexes membranaires multi-protéique en présence de multiples ligands[471], capside virale[472], *etc.*).

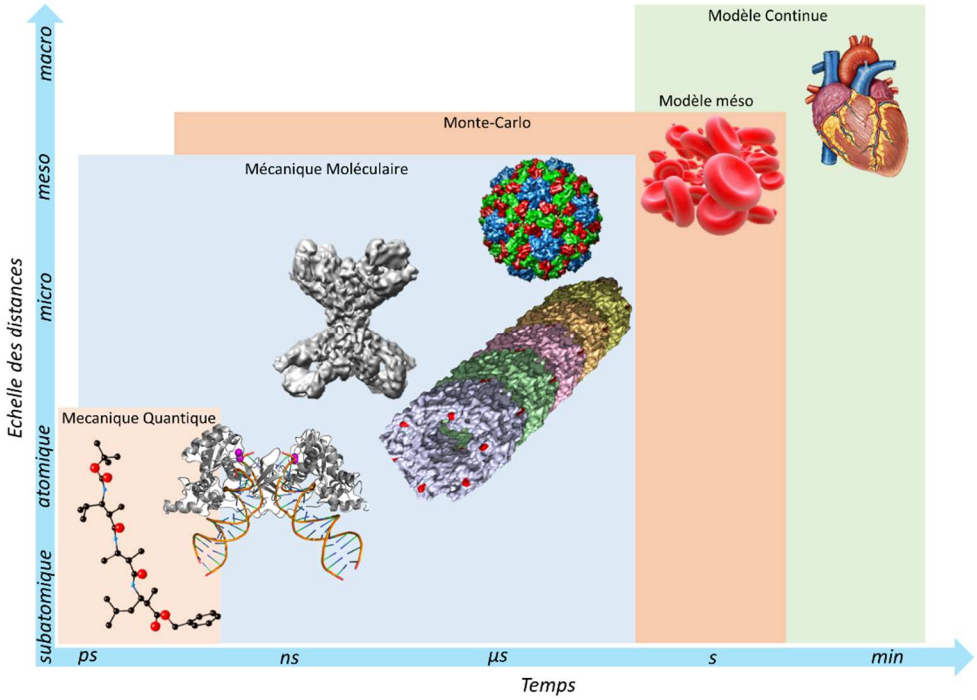


Figure 15 : Les processus moléculaires biologiques et l'échelle des temps associée.

Différents paramètres sont à prendre en compte lorsque l'on souhaite effectuer une **simulation de dynamique moléculaire** d'une macromolécule. Le processus étudié, la taille du système ainsi que les ressources en temps de calcul disponibles vont orienter le choix de la durée des simulations ainsi que de la méthodologie. Trois principales méthodologies – mécanique quantique (QM), mécanique moléculaire (MM), simulations gros grains (GG) – et leur dérivés – méthodes hybrides QM/MM et MM/GG – coexistent et sont employées en fonction des besoins, mais toutes reposent sur un principe commun : les mouvements atomiques sont dirigés par les forces qui s'exercent sur les atomes. Un champ de forces décrit ainsi les interactions que chaque atome établi avec son environnement, et permet ainsi de décrire la physique du système à chaque instant de la simulation. Ces champs de forces ont plusieurs origines : (1) une simulation de dynamique moléculaire dite *ab initio* ou quantique utilisera un potentiel énergétique en se basant sur les principes quantiques, prenant en compte la structure électronique des atomes du système ; (2) les simulations de dynamique moléculaire dites « classiques » utiliseront un champ de forces « empirique », dont les paramètres sont dérivés des bases de données structurales (PDB). Les méthodes classiques présentent comme avantage d'être moins couteuses en temps de calcul, mais ont les désavantages de ne pouvoir représenter des changements de topologie (formation/rupture de liaison covalente) et d'être moins précises car ne reposant pas sur la structure électronique des atomes. Le choix de la méthode à utiliser dépendra ainsi des caractéristiques de celles-ci : la modélisation de réactions chimiques ne peut ainsi être réalisée par des simulations de dynamique dite « classique ». À l'inverse, des

phénomènes longs (plusieurs nanosecondes et au-delà) sur de systèmes comportant plusieurs milliers d'atomes ne peuvent utiliser la dynamique quantique car trop coûteuse en temps de calcul.

Afin de répondre aux limitations explicitées ci-dessus (temps de simulations accessibles, coût des calculs, *etc.*), de nouvelles approches ont été introduites. Les deux approches les plus notables sont les simulations hybrides QM/MM et les simulations gros grains. Dans la première, le système est représenté pour une partie en mécanique quantique (QM) et pour l'autre en mécanique moléculaire/classique (MM). En limitant la description quantique à une petite partie du système, le temps de calcul est réduit considérablement, bien que supérieur à une simulation en mécanique classique, tout en profitant des avantages offerts par la mécanique quantique au niveau de la région d'intérêt (le site de liaison ou le site actif par exemple). La simulation en gros grains est à l'inverse une approche représentant des groupes d'atomes par une unique particule, un gros grain. La taille du système (comprendre le nombre de particules constituant le système) étant réduit, des simulations de très grands systèmes pendant de très longs temps sont ainsi accessibles même si on perd la description tout-atome des systèmes. Enfin, l'interaction entre le code du logiciel de simulation de dynamique moléculaire et l'architecture du cluster de calcul peut être optimisée : la machine Anton[473] peut générer des simulations de l'ordre de la milliseconde[474] en un temps d'éxécution inférieur à un cluster de calcul traditionnel.

2. Dynamique Moléculaire en mécanique classique

a) Les équations du mouvement

La représentation de la dynamique en mécanique classique est la plus courante dans le champ de la biologie computationnelle car la plupart des processus étudiés ne comportent pas d'évènements impliquant un changement de la connectivité des atomes (réactions enzymatiques) et qu'une description à l'échelle atomique est néanmoins nécessaire à la l'analyse fine des modifications structurales et dynamiques. Les simulations de dynamique moléculaire résolvent, pour l'ensemble des *N* atomes du système, les équations du mouvement de Newton :

$$F_i = m_i \frac{\partial^2 r_i}{\partial t^2} \, , i = 1 \dots N \qquad \textbf{Équation 1}$$

où F_i sont les forces s'exerçant sur l'atome i, m_i la masse de l'atome i, $r_i = (x_i, y_i, z_i)$ les coordonnées de l'atome i et t le temps.

Les forces sur les atomes à un instant t sont calculées à partir de la dérivée du potentiel d'énergie potentielle $V(r_1, r_2, \dots, r_N)$:

$$F_i(t) = -\frac{\partial V(r_1(t), r_2(t), \dots, r_N(t))}{\partial r_i(t)} \qquad \textbf{Équation 2}$$

Ces équations sont résolues simultanément de manière itérative pour un pas de temps δt, appelé par conséquent « pas d'intégration ». Chaque itération génère un nouveau jeu de

coordonnées qui, mises bout à bout, constituent une trajectoire de dynamique moléculaire. Les systèmes s'équilibrent et atteignent ainsi un état d'équilibre : ces simulations sont ainsi dites « à l'équilibre ».

Le pas d'intégration représente l'intervalle de temps séparant deux évaluations successives de la fonction d'énergie. Sa valeur doit être suffisamment petite pour ne pas discrétiser certaines quantités mesurables. Dans les faits, cela correspond à choisir un pas d'intégration inférieur au mouvement de plus haute fréquence du système (théorème de Nyquist-Shannon[475]) : la fréquence d'un mouvement étant liée à la masse des particules impliquées dans ce mouvement vibratoire, la plus haute fréquence des systèmes biologiques est liée à l'élongation des liaisons covalentes impliquant des atomes d'hydrogène (liaisons C–H, O–H ou N–H). La fréquence vibratoire de ces mouvements est supérieure à 3000 cm^{-1} à 310K, d'où la nécessité d'utiliser un pas d'intégration de l'ordre de 1 femtoseconde (1 fs = 10^{-15} s).

L'intégration des pas de temps peut se faire par plusieurs méthodes différentes, nous présenterons ici l'algorithme dit « *leapfrog* », ou « saut de grenouille »[476]. Cet algorithme est une méthode de résolution numérique de l'équation différentielle (équation 1), en utilisant les positions atomiques r au temps t et les vitesses atomiques v au temps $-\frac{1}{2}\delta t$. Les positions et vitesses sont alors calculées grâce aux équations suivantes :

$$r(t+\delta t) = r(t) + \delta t \cdot v(t+\tfrac{1}{2}\delta t) \qquad \textbf{Équation 3}$$

$$v\left(t+\tfrac{1}{2}\delta t\right) = v\left(t-\tfrac{1}{2}\delta t\right) + \frac{\delta t}{m}F(t) \qquad \textbf{Équation 4}$$

Cet algorithme calcule les vitesses des atomes de manière explicite (à l'inverse de l'algorithme de Verlet[477] par exemple), bien qu'à l'instant $t+\delta t$. Les vitesses à l'instant t peuvent être approximées grâce à la relation :

$$v(t) = \tfrac{1}{2}\left[v\left(t-\tfrac{1}{2}\delta t\right) + v\left(t+\tfrac{1}{2}\delta t\right)\right]. \qquad \textbf{Équation 5}$$

Afin de démarrer une dynamique moléculaire, un jeu de vitesses initiales doit être fourni afin d'amorcer la dynamique du système. Si aucune donnée ne permet d'obtenir les vitesses initiales, un jeu de vitesses à $t = t_0 - \frac{1}{2}\delta t$ sera généré automatiquement en suivant une distribution aléatoire de Boltzmann à une température donnée :

$$p(v_i) = \sqrt{\frac{m_i}{2\pi k_B T}}\, exp\left(-\frac{m_i v_i^2}{2k_B T}\right), \qquad \textbf{Équation 6}$$

où m_i et v_i la masse et la vitesse de l'atome i, k_B la constante de Boltzmann, $k_b = 8{,}314510 \cdot 10^{-3}\mathrm{kJ\ mol^{-1}K^{-1}}$ et T la température.

L'énergie cinétique du système Ec est obtenue à partir des vitesses calculées grâce à l'équation 6 et doit respecter l'égalité :

$$E_c = \sum_{i=1}^{N} \frac{1}{2} m_i v_i^2 = \frac{1}{2} N_{DL} k_B T, \quad \text{Équation 7}$$

où $N_{DL} = 3N - N_c - N_{com}$, N est le nombre d'atomes du système, N_c est le nombre de contraintes appliquées au système et $\begin{cases} N_{com} = 3 \\ N_{com} = 6 \end{cases}$ en fonction des mouvements de translation et/ou de rotation du système qui sont retirés du mouvement. Une fois les vitesses initiales connues, l'intégration des formules des équations 3 et 4 est possible.

b) Champ de forces

La fonction d'énergie utilisée dans l'équation 2 est décrite par le champ de force, qui comprend des termes reliés à la fois aux interactions liantes (les liaisons covalentes) et non-liantes (interactions électrostatiques et de van der Waals). Les composants détaillés du champ de forces sont donc les suivants :

$$E_{totale} = E_{liantes} + E_{non-liantes}, \quad \text{Équation 8}$$

avec $\begin{cases} E_{liantes} = E_{élongation} + E_{angle} + E_{dièdre} \\ E_{non-liantes} = E_{électrostatique} + E_{van\ der\ Waals} \end{cases}$. Équation 9

L'énergie d'élongation correspond à la variation de la longueur des liaisons covalentes, et est représenté par une fonction de potentiel harmonique générique de type

$$E_{élongation} = \sum k(r_{ij} - r_{ij0})^2, \quad \text{Équation 10}$$

où k est la constante de force de liaison, r_{ij} est la longueur instantanée de la liaison entre les atomes *i* et *j*, et r_{ij0} est la longueur de la liaison de référence. Les variations de l'angle formé par trois atomes sont également représentées par un potentiel harmonique :

$$E_{angle} = \sum k\,(\theta_{ijk} - \theta_{ijk0})^2, \quad \text{Équation 11}$$

avec k la constante de force angulaire, θ_{ijk} l'angle formé par les atomes *i*, *j* et *k* à l'instant *t*, et θ_{ijk0} l'angle *i-j-k* de référence. Les angles dièdres correspondent à la rotation de deux groupements autour d'une liaison et implique donc quatre atomes, *i, j, k* et *l*. L'angle dièdre ϕ autour de la liaison *j-k* est l'angle formé par les deux plans *i-j-k* et *j-k-l*. L'énergie des angles dièdres est donc décrite par la fonction générique

$$E_{dièdre} = \sum \sum_i k_\phi (1 + \cos(n\phi - \phi_s)), \quad \text{Équation 12}$$

où k_ϕ est la constante de torsion, n est la périodicité de la rotation et ϕ_s est l'angle de phase.

Les interactions non-liantes sont également représentées par les fonctions suivantes. Pour chaque paire d'atomes i et j séparés par une distance r, les interactions de van der Waals sont décrites par le potentiel de Lennard-Jones :

$$E_{van\ der\ Waals}(r_{ij}) = 4\varepsilon^0 \left[\left(\frac{r_{ij}^0}{r_{ij}} \right)^{12} - \left(\frac{r_{ij}^0}{r_{ij}} \right)^{6} \right], \qquad \textbf{Équation 13}$$

où r_{ij}^0 est la distance pour laquelle les forces attractives et répulsives s'annulent et ε^0 l'opposé de la valeur du potentiel au point r_{ij}^0 . Enfin les interactions électrostatiques entre les charges électriques q_i et q_j des atomes i et j, respectivement, sont données par :

$$E_{électrostatique} = \sum \frac{1}{4\pi\varepsilon_0} \frac{q_i q_j}{\varepsilon_r r_{ij}}, \qquad \textbf{Équation 14}$$

où ε_0 est la permissivité du vide et ε_r la permissivité du milieu.

Cette forme générique est partagée par de nombreux champ de forces, mais d'autres termes peuvent être ajoutés afin de mieux décrire les propriétés physiques du système, comme notamment les angles dièdres impropres afin de maintenir certains groupes (les cycles aromatiques par exemple) planaires. Chaque champ de forces est divergeant par les paramètres associés aux différentes fonctions (ex : la constante de force angulaire pour les mêmes atomes) et par ces termes supplémentaires (termes de polarisation, s'ils sont présents). Les paramètres des champs de forces ont deux origines possibles : pour les molécules disposant de larges données (typiquement les acides aminés et les acides nucléiques), les paramètres vont être extraits des bases de données afin de générer des champs de forces dits « empiriques ». Dans le cas contraire (ligands), les paramètres vont être dérivés de calculs de mécanique quantique.

Le calcul des forces non-liantes – électrostatiques (équation 13) et de van der Waals (équation 14) – se base sur la notion de paire d'atomes. L'inconvénient de cette méthode est que le nombre d'interactions varie avec N^2. Pour limiter l'impact du calcul de ces forces, une limite de distances est appliquée et seuls les atomes séparés par une distance inférieure à cette limite seront pris en compte pour le calcul des interactions non-liantes. Cette valeur seuil doit être inférieure à la moitié de la plus petite longueur de la boîte de simulation pour éviter l'interaction de particules avec l'une de leur image périodique. Des approches permettent ensuite de limiter l'effet de la troncation, qui implique dans la forme décrite ci-dessus une discontinuité dans le champ de forces : le potentiel peut être diminué pour atteindre la valeur

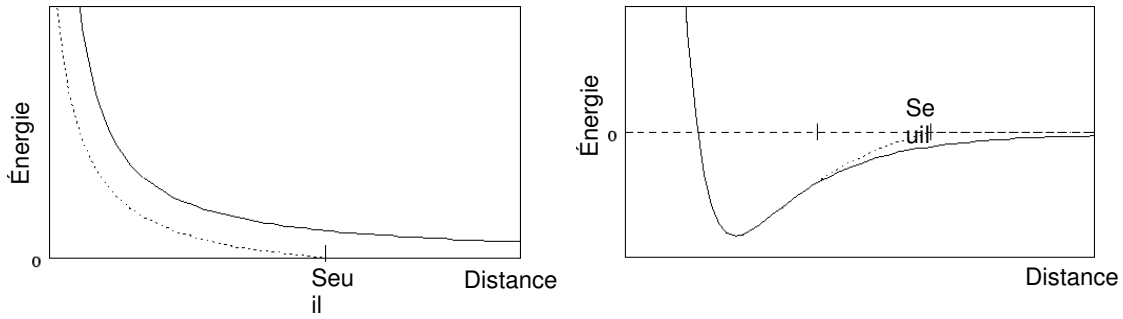


Figure 16: **Troncation des interactions non-liantes.** À gauche, l'application d'un "*shift*" change le potentiel, qui est nul pour la valeur seuil. À droite, le potentiel est décalé à partir d'une limite pour devenir nul à la valeur seuil. Figures issues du site du logiciel *NAMD* : http://www.ks.uiuc.edu/Research/namd/2.9/ug/node23.html

de zéro au niveau de la valeur seuil (« *shift* »), ou un modificateur est appliqué au potentiel à partir d'une distance pour atteindre la valeur de zéro au niveau de la valeur seuil (« *switch* ») (*cf.* Figure 16).

Enfin, pour les interactions longue distance, l'algorithme *Particle Mesh Ewald* (PME) est généralement utilisé, appliqué pour la première fois au calcul du potentiel électrostatique en 1993[478]. La sommation de l'énergie électrostatique d'un système et de ses images périodiques est lente à converger. Plus particulièrement, les interactions longue-distance sont les plus lentes à converger ; la sommation d'Ewald[479] propose de changer l'espace de calcul de ces interactions de l'espace réel vers un espace de Fourier, dont la convergence sera plus rapide. L'approche PME améiore les perfomances de la sommation d'Ewald en utilisant une grille et une transformation de Fourier rapide pour le calcul des interactions longue-distance. La complexité du calcul varie selon N•log(N) au lieu de N^2 pour la sommation d'Ewald, d'où un gain de temps considérable.

Le contrôle de la température est un élément important puisque la température d'un système est corrélée à son énergie cinétique et donc à la quantité de mouvements observée (*cf.* équation 7). De même, la pression est reliée au carré de la vitesse des particules *via* la relation 15:

$$P = \frac{1}{3}\frac{N}{V}mv^2. \qquad \textbf{Équation 15}$$

Le maintien de ces deux variables d'état (température et pression) à des niveaux constants au cours de la simulation correspond aux conditions physiologiques des protéines. Combiné à un nombre constant de particules, nous obtenons un ensemble isotherme-isobare, *NPT*. L'utilisation de l'ensemble microcanonique (*NVE*, nombre d'atomes constant, volume et énergie constants) ou canonique (*NVT*, nombre de particules constant, volume et température constants) peut également être envisagé ; cependant les erreurs liées aux approximations (résolution des équations du mouvement, *etc.*) engendrent des variations de température et de pression au cours du temps. Afin d'éviter ces dérives, le coulage à la température et à la pression sont couplés à des bains afin de maintenir ces valeurs constantes généralement réalisé grâce à la connexion du système avec des bains.

3. Les limites de la mécanique classique

Les simulations de dynamique moléculaire par mécanique classique reposent sur des théories physiques bien définies, mais qui présentent néanmoins certaines limites. La confrontation des données obtenues est un moyen de contrôle quant à la validité et la précision des simulations réalisées. Cependant, ces données ne sont pas forcément disponibles, et surtout elles ne sont pas établies aux mêmes échelles de grandeur de temps (heures *versus* nanosecondes) et de taille (échelle macroscopique *versus* atomistique).

Les trajectoires de dynamique moléculaire sont toujours analysées sous l'hypothèse ergodique : la moyenne d'une grandeur mesurée sur les conformations générées est égale à la moyenne de cette grandeur au cours du temps. En statistique mécanique, cette hypothèse permet de s'affranchir partiellement de la nécessité d'explorer l'ensemble de l'espace conformationnel (l'ensemble des conformations que peut adopter une molécule), si un nombre suffisant de conformations représentatives est généré. Dans la pratique, parcourir l'ensemble de cet espace conformationnel par simulations de dynamique moléculaire n'est pas réalisable, car il est immense. Générer plusieurs trajectoires permet de calculer les valeurs d'intérêt en générant d'autres conformations représentatives de l'ensemble en utilisant des vitesses initiales différentes, les sous-espaces conformationnels explorés dans chaque dynamique diffèrent partiellement. Certaines approches computationnelles permettent de générer un plus grand nombre de conformations en orientant/limitant les déplacements du système.

Parmi celles-ci, la **méta-dynamique** est utilisée lorsqu'un obstacle énergétique important empêche l'exploration du paysage conformationnel[480]. Un potentiel est ajouté au paysage énergétique afin de limiter le retour du système aux points déjà exploré : les puits énergétiques vont ainsi être comblés progressivement et la barrière énergétique va devenir accessible. Les **simulations de dynamique dirigée**, à partir d'une structure de départ et d'une structure cible, vont guider linéairement l'évolution du système de l'un à l'autre en appliquant un potentiel harmonique aux atomes[481]. Si cette approche permet de franchir rapidement des obstacles énergétiques, elle peut également déformer la structure lors de réarrangements majeurs et ne suit pas nécessairement le parcours de plus faible énergie. Les simulations de dynamiques guidées appliquent également des contraintes, mais elles l'appliquent au centre de masse du système et non à des atomes. Ainsi, les déformations de structures sont moindres, mais le système ne suit pas forcément les variations du paysage énergétique. D'autres méthodes ont été développées afin de répondre à ces limites, mais ce champ d'étude reste à ce jour très ouvert[482,483].

L'utilisation de la mécanique classique pour décrire le système et intégrer les équations de Newton est suffisamment précise pour la plupart des atomes à température ambiante, mais peut présenter certaines limitations dans le cas de transfert des protons, dont la dynamique est essentiellement dominée par des effets quantiques. Les liaisons hydrogènes sont un exemple de ces effets à prédominance quantique. En pratique, tous les mouvements vibratoires rapides (dont la fréquence est supérieure à 200 cm^{-1}) peuvent potentiellement se comporter d'une

manière inappropriée. Appliquer des contraintes sur ce type de mouvements (notamment les mouvements vibratoires impliquant des atomes d'hydrogène) va limiter ces effets, en plus de permettre l'augmentation du pas d'intégration, donc du temps de simulation pour un temps de calcul donné. D'autres types d'interactions peuvent être biaisés, il s'agit des liaisons non-liantes. Comme décrit dans le paragraphe précédent, ce type de forces est tronqué à partir d'une distance seuil. De plus, les champs de force ne modélisent pas les effets de polarisabilité qui prennent place dans les molécules. Dans la pratique, ces effets ne sont pas (trop) mal traités par les champs de forces, ce qui explique que nous les utilisions toujours.

Les paramètres des champs de forces sont issues des données structurales, par nature fortement hétérogèes. Les conditions expérimentales (température notamment) qui ont permis de générer ces données peuvent fortement fluctuer, et diffèrent des conditions de simulation. Notamment, certaines simuations introduisent des variations de température amples (recuit simulé par exemple) ; les paramètres du champ de forces peuvent ainsi se révéler de faible qualité et amener des distorsions de la structure atomique des molécules simulées. Ainsi, si l'utilisation de champ de forces empiriques démontre une robustesse croissante, ces outils restent perfectibles et sont enrichis ou corrigés régulièrement.

E. Analyse des Modes Normaux

Les modes normaux sont utilisés pour caractériser les mouvements oscillatoires d'un système, au voisinage de son état d'équilibre. Ils permettent de décrire les mouvements vibratoires lorsqu'un objet est excité à des fréquences particulières. L'analyse des **modes normaux** n'est pas réservée aux objets biologiques : elle a été développée et appliquée à d'autres systèmes, pour caractériser tous les modes vibratoires. Un mode normal est le mouvement associé à un mouvement vibratoire intrinsèque à une fréquence donnée lorsque le système est perturbé à proximité d'un état stable. La superposition des modes normaux est ainsi capable de reproduire le mouvement général d'un système, alors représenté comme un assemblage d'oscillateurs harmoniques. Si la structure étudiée est trop éloignée de l'état d'équilibre, des effets non-linéaires apparaissent et les modes ne sont alors plus normaux, donc l'excitation d'un mode sera associée à l'excitation d'autres modes. En calculant les modes normaux pour un système protéique, les modes de plus basse fréquence représenteront les mouvements les plus lents et les plus amples de la protéine, généralement de l'ordre de la milliseconde. L'analyse des modes normaux permet ainsi de compléter la description des mouvements observés par dynamique moléculaire, limitée à l'observation de phénomènes de l'ordre de la microseconde. Ces modes de plus basse fréquence possèdent plusieurs propriétés intéressantes : ils sont spécifiques d'un système donné, sont généralement associés à la fonction de la protéine, impliquent de manière coordonnée un grand nombre d'atomes (jusqu'à constituer des domaines) et caractérisent des mouvements amples (jusqu'à plusieurs angströms)[484]. A l'inverse, les modes de plus haute fréquence décrivent des mouvements de moindre amplitude et impliquant un nombre limité d'atomes. Enfin, le calcul des modes normaux, même si il ignore *de facto* les mouvements anharmoniques, est nettement moins

consommateur de temps de calcul que les simulations de dynamique moléculaire, ce qui en fait une approche intéressante dans l'étude des phénomènes biologiques impliquant de larges réarrangements structuraux.

Le calcul des modes normaux s'appuie sur l'approximation harmonique localisée de la surface d'énergie potentielle du système (*cf.* Figure 17). Au point d'énergie minimale q^0, la fonction d'énergie peut être approximée par un développement en série de Taylor:

$$V(q) = V(q^0) + \sum_i \left(\frac{\partial V}{\partial q_i}\right)^0 (q_i - q_i^0) + \frac{1}{2}\sum_{i,j} \left(\frac{\partial^2 V}{\partial q_i \partial q_j}\right)^0 (q_i - q_i^0)\left(q_j - q_j^0\right) \ldots \text{Équation 16}$$

où qi est la i-éme coordonnée, V_0 est l'énergie potentielle de la structure dont on veut calculer les modes de vibration ; l'indice « 0 » est associé aux coordonnées de la conformation considérée.

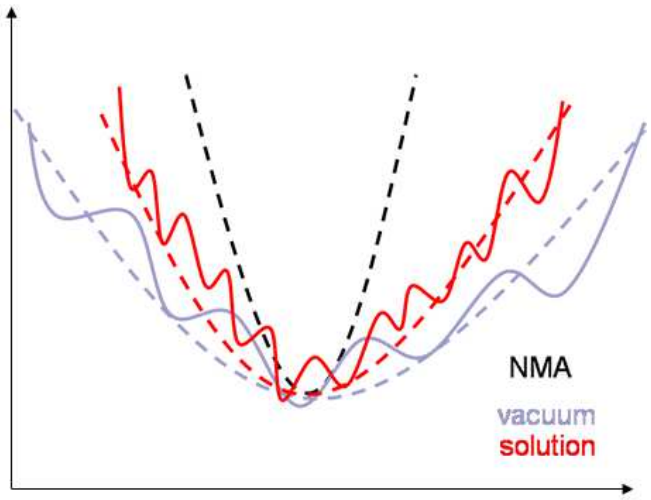


Figure 17 : Approximation quadratique de la surface d'énergie potentielle par les modes normaux (noir), comparée aux surfaces d'énergies de simulations de dynamique moléculaire dans le vide (gris) ou en solvent explicite (rouge). Issu du site internet : http://mmb.irbbarcelona.org/FlexServ/help/book.php

Au point d'équilibre, les deux premiers termes de l'équation 16 sont nuls. Si les déplacements atomiques sont suffisamment petits, on peut négliger les termes d'ordre supérieur du développement en série. On obtient que l'énergie potentielle peut donc être approximée par :

$$V(q) = \frac{1}{2}\sum_{i,j} \left(\frac{\partial^2 V}{\partial q_i \partial q_j}\right)^0 (q_i - q_i^0)\left(q_j - q_j^0\right)$$

$$= \frac{1}{2}\sum_{i,j} (q_i - q_i^0) H_{ij} \left(q_j - q_j^0\right)$$

$$= \frac{1}{2}\Delta q^T H \Delta q \,, \qquad \textbf{Équation 17}$$

où H est la matrice hessienne obtenue à partir des dérivées secondes par rapport aux déplacements atomiques :

$$H_{ij} = \left(\frac{\partial^2 V}{\partial q_i \partial q_j}\right)^0$$ **Équation 18**

La matrice hessienne H, qui est réelle, symétrique par construction, est composée de $N \times N$ sous-matrices (N étant le nombre de particules qui composent le système) qui décrivent la contribution de chaque paire de particules à la fonction d'énergie. Les valeurs propres de la matrice hessienne seront positives, à l'exception de six valeurs propres qui décrivent les mouvements de rotation et translation de la molécule. L'équation 18 peut être réécrite sous la forme du problème aux valeurs propres généralisé :

$$A^T H A = \lambda\,,$$ **Équation 19**

où chaque vecteur propre A_k représente une mode normal, et chaque valeur propre λ_k est lié au coût énergétique du déplacement le long du mode k, $k = 1 \ldots 3N$. De plus, les valeurs propres sont également liées au carré des fréquences des modes normaux par la relation $\lambda_k = \omega_k^2$, où ω_k est la fréquence vibratoire du mode k [485].

Une limite immédiate des modes normaux est l'approximation quadratique d'un minimum local : les modes déterminés ne sont ainsi valables qu'à proximité de la structure à l'équilibre. Les mouvements le long des modes doivent donc être regardée avec précaution, d'autant plus lorsqu'on s'éloigne de la structure à l'équilibre. De même, les modes normaux sont des déplacements instantanés et sont par définition tangents au mouvement à l'équilibre. La structure équilibrée contient des contraintes internes (longueur des liaisons, angles, *etc.*) qui doivent être réévaluées lorsqu'on se déplace le long des modes normaux, et une nouvelle structure doit être déterminée à un nouveau minimum local. La recherche de chemins de transition complexes nécessite donc en général un processus itératif de déplacement le long d'un mode suivi par une minimisation énergétique suivie par le calcul de modes normaux[486], afin de limiter les biais potentiels. La structure utilisée pour les modes normaux étant un modèle non solvaté, les modes normaux ne prennent pas en compte l'effet du solvant : la fréquence des modes normaux (calculés à partir des valeurs propres) pour les modes lents est donc sous-estimés. Pour les mouvements conformationnels importants, l'échelle de temps des mouvements de large amplitude ne peut être assimilée à l'échelle de temps d'oscillateurs harmoniques libres. Enfin, une autre limitation importante de cette méthode est liée à la taille de la matrice hessienne, dont la diagonalisation est requise la réduction.

Plusieurs approches ont été proposées afin de diminuer les ressources nécessaires au calcul des modes normaux. Ces méthodes ont à la fois cherché à améliorer l'efficacité des différentes étapes de calcul[487–489] et à diminuer le nombre de degré de liberté.

Cette dernière approche à mener à l'élaboration de plusieurs modèles dits à gros grains, adoptant une description moins fine qu'une description tout-atome sans perdre trop d'information sur la dynamique du système. Deux principaux modèles sont à souligner. D'une part, le modèle en bloc introduit par Yves-Henri Sanejouand et collaborateurs[490,491] regroupe un ou plusieurs résidus en un seul bloc rigide, ne possédant que six degrés de liberté. Les modes

normaux calculés sont alors une combinaison linéaire des mouvements de rotation-translation des blocs. L'autre approche a été initiée par Monique Tirion[492], qui a remplacé les interactions entre chaque paire d'atomes par des ressorts harmoniques, si les atomes sont séparés d'une distance inférieure à une distance seuil. Les forces inter-atomiques sont alors modélisées par un potentiel de Hooke :

$$E(r_a, r_b) = \frac{k}{2}\left(|r_{a,b}| - |r^0_{a,b}|\right)^2, \qquad \text{Équation 20}$$

où k est la constante de force du ressort, $r_{a,b}$ et $r^0_{a,b}$ sont les distances instantanées et à l'équilibre entre les atomes a et b, respectivement. L'énergie totale du système devient alors :

$$V = \sum_{(a,b)} E(r_a, r_b). \qquad \text{Équation 21}$$

Cette étude a été le premier modèle en réseau élastique tout-atome, dont les liaisons et angles inter-atomiques étaient fixes. Très vite, d'autres modèles élastiques ont été décrits, où chaque acide aminé d'une protéine est représenté par une particule, sans considération du type d'atome ou de la masse du résidu. Ces modèles ont montré de très bons résultats pour générer des aperçus précis des propriétés dynamiques des macromolécules, notamment concernant les mouvements fonctionnels de grande amplitude dans le cas de grands systèmes[492]. Deux types de modèles à réseau élastique peuvent être envisagés : un modèle unidimensionnel proposé par Ivet Bahar, le modèle en réseau gaussien (GNM)[493], ou un modèle tridimensionnel inspiré de Tirion. La constante du ressort harmonique peuvent être inspiré d'un champ de force[484] ou varier en fonction de la connectivité des paires d'atomes, comme pour le modèle en réseau anisotrope (ANM)[494], une extension du modèle GNM.

Tous ces modèles élastiques présentent un avantage important : ils ne nécessitent pas une minimisation de la structure avant le calcul des modes normaux. L'étape de minimisation, qui peut être coûteuse en temps de calcul et produire des structures douteuses, est ainsi évitée ; une conformation pertinente est néanmoins requise pour obtenir des résultats exploitables. Ce type de représentation utilise peu de ressources, est adaptable à tous les niveaux de gros-grains et requiert peu de paramètres : le choix des constantes de force des ressorts n'ayant qu'une faible influence (les modes normaux sont davantage des propriétés intrinsèques de la forme tridimensionnelle de la protéine).

III. L'allostérie est un phénomène fondamental en biologie

Les organismes biologiques multicellulaires sont composés d'une immense variété de cellules, chacune ayant différents fonctions selon l'organe ou le tissu où elle réside. Pour le développement et le fonctionnement normal d'un organisme pluricellulaire, le comportement de chaque cellule doit être strictement régulé. Cette régulation est basée sur un système de communication complexe comprenant la signalisation entre organes à travers l'organisme,

entre les différents types de cellules et entre les molécules d'une même cellule. Les communications entre cellules distantes de l'organisme mettent en jeu des molécules messagers émises dans le milieu extérieur et captées par les cellules cibles. Ces molécules messagers sont souvent trop polaires et /ou de trop grande taille pour diffuser à travers les barrières lipidiques comme la membrane cellulaire. Elles doivent donc être reconnues à la surface de la cellule cible par des récepteurs transmembranaires, qui relaient ces signaux de l'extérieur vers l'intérieur de la cellule.

Par conséquent, les protéines de différent type sont au cœur de tous les processus intra- et intercellulaires dans tous les organismes vivants et établissent les liens entre génétique, biochimie et métabolisme. Afin d'assurer ces différents rôles, l'activité des protéines est strictement et finement régulée/modulée par des effecteurs qui peuvent se fixer/lier sur un site protéique tout en modulant l'activité d'un second site, distant du site de fixation d'effecteur – on parle **de régulation allostérique** et/ou de couplage de sites. Les effecteurs allostériques sont des molécules autres que les substrats et dont la fixation en des sites distincts des sites de fixation des substrats diminue l'activité de l'enzyme (inhibiteurs) ou l'augmente (activateurs).

Le développement d'un schéma général dans lequel inscrire l'allostérie et la caractérisation des mécanismes moléculaires liés à ces phénomènes constitue un défi majeur en biologie. L'allostérie a ainsi été décrite par Jacques Monod[495] comme le « *second secret de la vie* » après le code génétique, mais reste encore mal décrite, quantifiée et prédite à l'échelle atomique. Une description quantitative de l'allostérie est fondamentale pour la compréhension de tous les processus au-delà de l'échelle moléculaire.

Les modèles de la régulation allostérique ont été initialement basés sur l'observation de structures cristallographiques figées[496]. Puis, le concept d'allostérie a été développé à partir de l'observation de structures oligomériques qui oscillaient entre deux états thermodynamiquement stables[497]. L'amélioration des techniques expérimentales a mis en évidence des phénomènes plus complexes qui nécessitaient l'élaboration de nouveaux modèles. Ces modèles, dynamiques par essence, mettent en jeu des mécanismes moléculaires spécifiques, qui peuvent être quantifiés, et impliquent des modifications des vastes réseaux inter-résidus se traduisant par des transitions conformationnelles dynamiques de la protéine. Ces transitions allostériques sont initiées par des évènements moléculaires tels que la fixation d'un ligand ou d'une biomolécule (protéine, nucléotide ou lipide), de la lumière, des modifications post-traduction (phosphorylation, glycosylation, …) ou des changements de l'environnement cellulaire. Élucider les mécanismes allostériques serait une étape clé qui permettrait de prédire les sites allostériques à la surface d'une protéine ou les mutations de résidus pouvant conduire à des effets allostériques, de décrire les chemins de propagation du signal, et de participer à la découverte de modulateur de l'activité biologique des protéines.

Afin de comprendre comment les théories actuelles ont été élaborées et explorées, le modèle initial de l'allostérie est présenté brièvement. Un résumé de la perception actuelle de l'allostérie sera ensuite exposé, suivi d'illustrations par des études expérimentales portant sur

l'allostérie. Enfin, l'exploration computationnelle visant la description des phénomènes allostériques et l'apport de ces phénomènes dans l'arsenal thérapeutique seront discutés.

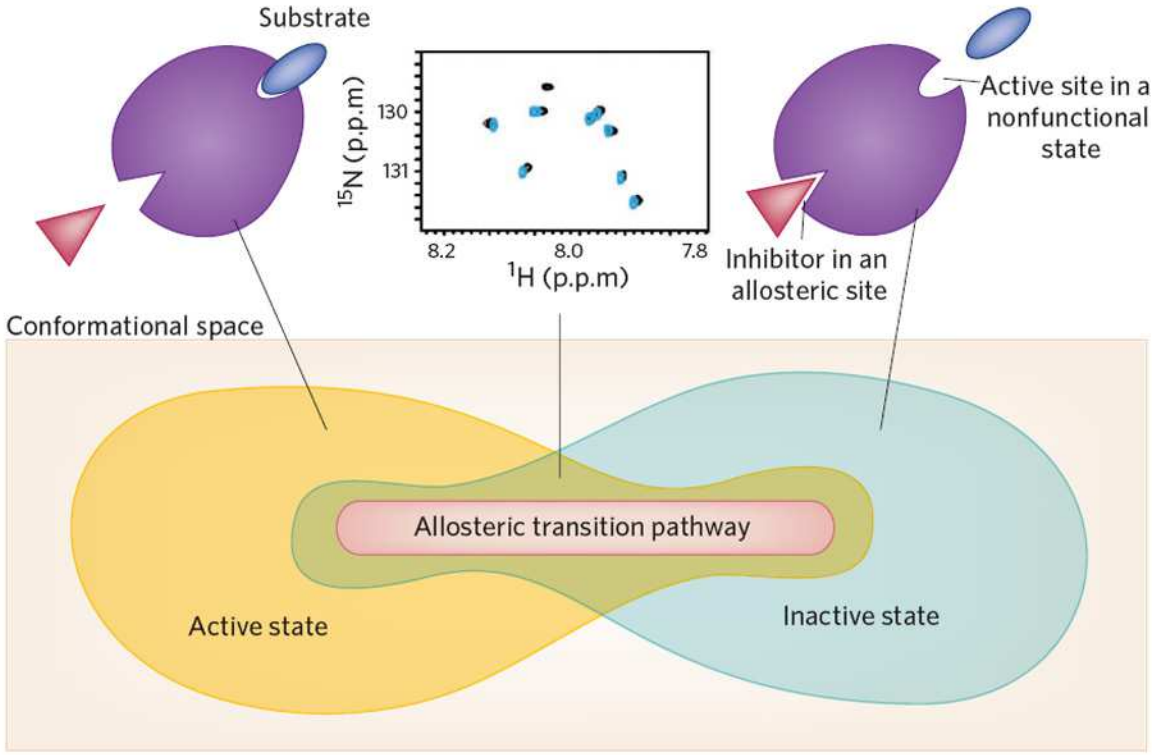


Figure 18: **Une protéine coexiste dans plusieurs états conformationnels distincts et reliés.** Le passage d'un état à l'autre, la transition allostérique, est lié à la présence ou l'absence de modulateurs qui se fixent sur des sites distants des sites de fixation des substrats. Des chemins de communication allostériques assurent la transmission du signal entre ces sites. Figure reproduite avec la permission des auteurs[498].

A. Découverte et description du phénomène allostérique

A début du XXème siècle (1904), Christian Bohr a étudié la liaison du dioxygène par l'hémoglobine dans différentes conditions. Le graphe de la saturation de l'hémoglobine en fonction de la pression partielle de l'oxygène a une forme sigmoïdale. Cela indique que plus le nombre de molécules de dioxygène liées est grand, plus l'affinité de l'hémoglobine augmente – jusqu'à ce que tous les sites soient occupés[499]. De plus, Bohr a noté que l'augmentation de la pression partielle en CO_2 déplace la courbe sur la droite - donc que les concentrations élevés en CO_2 rendent difficiles la fixation de dioxygène sur hémoglobine – l'effet Bohr, exprimé par A.V. Hill en formule analytique pour décrire la fixation coopérative à des protéines ayant des sites de liaison multiples[500].

Le terme "allostérique" vient des deux mots Grecs "*allos*" et "*stereos*" qui signifient : "*une autre forme*" ou "*un autre solide*". Il peut donc s'entendre comme "*une autre conformation*". Il a été employé par J. Monod pour expliquer le mode d'action des molécules

d'hémoglobine. Chacune des quatre sous-unités globine existe sous deux états en équilibre, un état tendu « T » peu affin pour le dioxygène et un état relâché « R » très affin pour le dioxygène (cf. Figure 19). L'état relâché d'une sous-unité expose un site de fixation pour un petit métabolite, le biphosphoglycérate, en un site distinct du site de fixation du dioxygène (la molécule d'hème). L'arrivée du biphosphoglycérate induit des modifications structurales de la globine qui perturbent les liaisons faibles qu'elle établit avec les autres sous-unités. Cet effet favorise la transition des autres sous-unités de l'état « T » à l'état « R », et augmente l'affinité totale de l'hémoglobine pour le dioxygène. Des phénomènes de régulation similaires ont ensuite été mis en évidence chez d'autres protéines multimériques, comme les récepteurs à l'acétylcholine[501]. À partir de ces observations, Monod, Wyman et Changeux ont proposé un modèle allostérique connu aujourd'hui sous le nom de « modèle MWC »[502]. Selon ce modèle, les protéines allostériques sont des protéines multimériques présentant une symétrie axiale et existantes dans un nombre discret fini d'états actifs ou inactifs réversibles. Le passage d'un état à l'autre est favorisé ou défavorisé par la liaison du ou des effecteur(s) allostérique(s), et la liaison du/des effecteur(s) est conditionnée à l'état de la sous-unité à laquelle il(s) se fixe(nt).

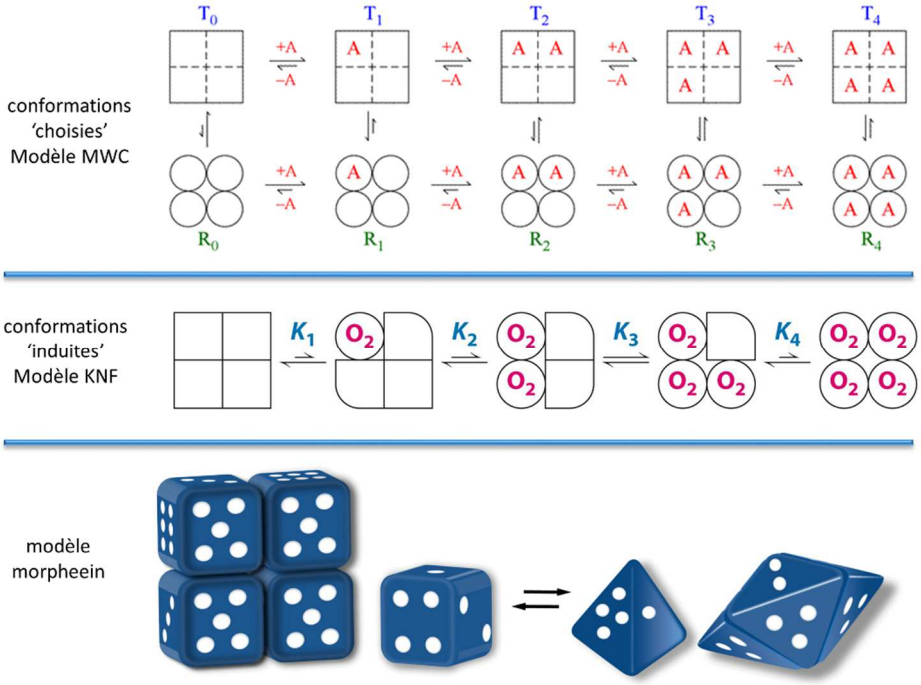


Figure 19: Représentation schématique des plusieurs modèles allostériques.

À ce modèle d'allostérie par conformations sélectionnées, l'équipe de Koshland a opposé un modèle d'allostérie par conformations induites, selon lequel la transition conformationnelle d'une sous-unité induite par la liaison de l'effecteur modifie l'affinité des sous-unités voisines pour leurs effecteurs, ce qui entraine ou réprime leur transition conformationnelle[503] (*cf.* Figure 19). Ce modèle séquentiel, connu sous le nom de modèle

Koshland-Nemethu-Filmer (KNF), est une généralisation de la théorie de Koshland sur les effets induits sur une enzyme par son substrat.

Ces deux modèles, MWC et KNF, découlent de l'observation de structures statiques (*i.e.*, cristallographiques) de protéines qui présentent alors deux état bien définis dont l'activité diffère, et adoptent soit l'un, soit l'autre via l'action d'une molécule effectrice sur un site allostérique. Les exemples représentatifs d'observation expérimentale (par étude cristallographique) des effets allostériques prédits par modelés sont (i) l'interaction allostérique entre un ligand et le site active de rhodopsine - un récepteur couplé aux protéines G (GPCR)[504] et (ii) la régulation allostérique de canal dans le récepteur membranaire AMPA[505]. De manière intéressante, ces deux modèles validés permettaient de décrire l'allostérie et reposaient sur les différences structurales observées entre différents états d'une même protéine.

Ces modèles, MWC et KNF, sont étroitement liés à l'observation des résultats expérimentaux et apportent une explication à ces résultats sans en détailler les mécanismes, notamment la manière dont la structure protéique transmet l'information allostérique entre les sites de liaisons des effecteurs et les sites fonctionnels.

B. Evolution des théories sur l'allostérie

De nouveaux modèles ont ensuite cherché à combler ce manque en faisant appel à des termes structuraux (*via* l'étude des structures à haute résolution) pour expliquer les phénomènes allostériques. Ce point de vue structural des mécanismes allostériques a ainsi occupé le devant de la scène, mais ces modèles élaborés à partir uniquement des structures étaient incomplets. Des approches thermodynamiques sont ainsi venues se greffer, impliquant de nouveaux mécanismes intra-protéine, mais également d'autres vecteurs tels que les molécules d'eau dans la régulation allostérique de l'hémoglobine[506]. Par ailleurs, Cooper et Dryden ont proposé un modèle général dans lequel l'allostérie se manifesterait par des changements dans la distribution des conformations, ce qui implique une contribution entropique à l'allostérie[507]. Cette approche statistique purement entropique a ainsi montré qu'une communication entre des sites distants est possible sans changements structuraux notables du site de liaison du ligand naturel sous l'influence de la liaison d'un effecteur allostérique, et que l'énergie libre associée à ces interactions est de l'ordre de quelques kilocalories par mole. Enfin, plus récemment, la découverte de phénomènes allostériques dans les protéines intrinsèquement désordonnées (*IDPs*, *Intrinsic Disordered Proteins*) a clairement démontré que la régulation allostérique est importante à la fois pour les domaines présentant une structure tertiaire, mais également pour les régions non-structurées qui présentent des fluctuations conformationnelles importantes.

Plus tard, (en 2005) Jaffe a formulé un modèle "dissociatif" concerté : un homo-oligomère peut exister sous plusieurs formes physiologiquement et fonctionnellement pertinentes en alternant entre différents assemblages quaternaires[508]. Les transitions entre ces différents assemblages alternatifs, appelés 'morpheein', impliquent la dissociation de

l'oligomère, un changement conformationnel de l'état dissocié puis un ré-assemblage avec un oligomère diffèrent. L'étape de dissociation des oligomères est l'élément qui différentie ce modèle des modèles MWC et KNF (*cf. Figure 19*). La découverte des 'morpheeins' a révélé un mécanisme inattendu pour cibler des enzymes ubiquitaire par des composés développés spécialement pour une espèce donnée. Un inhibiteur de ce type se lierait et stabiliserait l'état 'morpheein' inactif de l'enzyme, modifiant l'équilibre des formes vers cette dernière.

La description quantitative des mécanismes de régulation de l'allostérie de certaines protéines est basée sur ces différents modèles, mais la représentation généraliste et complète du phénomène constitue toujours un défi à relever. Dans ce cadre, il apparaît clairement que les facteurs qui permettent la transition entre les différentes conformations allostériques sont aussi importants que les changements structuraux lorsqu'ils existent. La description du paysage conformationnel d'une protéine constitue ainsi un enjeu : les différents états, dont le passage de l'un à l'autre est régulé par un large ensemble de conditions environnementales (présence d'un effecteur, pH, *etc.*) et internes (plasticité, réversibilité). Les études menées ont ainsi montré que les mécanismes moléculaires permettant ces transitions peuvent se traduire par des changements notables dans les structures de l'ensemble conformationnel d'une protéine, ou uniquement par de subtiles modifications structurales, voire l'absence de modification structurale. Cette dernière propriété a favorisé l'émergence d'une définition purement thermodynamique de l'allostérie, dans laquelle l'allostérie est contrôlée par l'enthalpie, l'entropie ou les deux[509].

En adoptant une approche énergétique, le repliement d'une protéine peut être décrit à partir du paysage énergétique conformationnel propre à cette protéine, et donc le couplage allostérique peut également l'être. Dans ce cadre, une conformation de faible énergie sera explorée plus souvent qu'une conformation de plus haute énergie. De plus, la « surface » d'énergie associée peut être lisse, favorisant ainsi l'exploration très libre du paysage et donc de multiples conformations, ou plus discrète, seuls un nombre limité de conformations étant alors accessibles. Les variations des énergies du paysage ou de couplage, qui sont en générales faibles, corroborent l'idée qu'à un instant *t* donné, il n'y a ainsi pas une unique conformation accessible, mais plusieurs conformations qui coexistent : un sous-ensemble de conformations va néanmoins dominer le paysage énergétique. La liaison d'un ligand ou la présence d'un effecteur allostérique à une conformation donnée va *de facto* remodeler le paysage énergétique autour d'un nouveau sous-ensemble conformationnel (*cf.* Figure 20). Présenté d'une manière différente, les mécanismes allostériques proposés sont plus statistiques que déterministes. Une nouvelle description de la protéine est ainsi apparue, dans laquelle chaque site de liaison est considéré individuellement comme pouvant prendre deux états (actif ou inactif), et interagir avec les autres[510,511]. Cette approche permet la description fine de la régulation allostérique de chaque sous-ensemble énergétique : la combinaison des différents sites de liaison dans chacun de leur état (actif ou inactif) et de la présence d'un ligand permet en effet de décrire tous les sous-ensembles de la protéine, et de les effets des effecteurs [509]. Couplés aux données expérimentales, certaines propriétés peuvent être énoncées[512].

Premièrement, une perturbation peut induire des effets différents, en fonction de l'équilibre du système au moment de la perturbation. Elle influence donc la fonction de la protéine mais également l'interprétation des changements d'entropie conformationnelle du système au cours de la transition allostérique, qui, pour une transition donnée, peut être soit augmentée, soit diminuée (*cf.* Figure 20). Deuxièmement, les protéines régulées de manière allostérique sépare les sites dont les fonctions diffèrent : les sites de liaison des effecteurs allostériques sont ainsi en général éloignés des sites fonctionnels. Si chaque site de liaison à un effecteur a une conformation active et une conformation inactive, la liaison d'un effecteur va contraindre ce site dans la forme active tandis que le site fonctionnel va être stabilisé ou déstabilisé. D'après cette théorie du paysage énergétique, si plusieurs sites de liaison existent, des sous-ensembles conformationnels existent, dans lesquels les sites de liaison sont couplé aux sites fonctionnels soit de manière positive, soit de manière négative. Le sens du couplage va alors dépendre de l'équilibre de l'ensemble conformationnel au moment de l'apparition de la perturbation, donc de l'équilibre entre forme active et forme inactive de chacun des sites, donc de la stabilité relative de chacun des états du système. Il déterminera également le pouvoir répresseur ou activateur de l'effecteur sur le système régulé (*cf.* Figure 20).

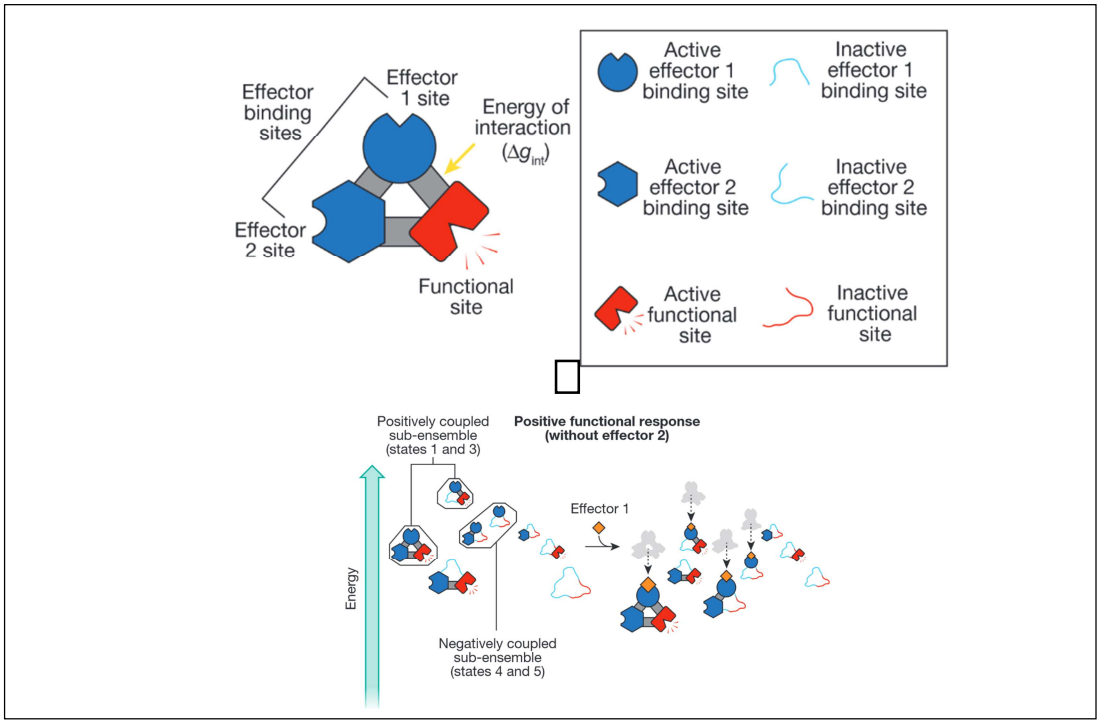


***Figure 20:* Le modèle à ensemble de l'allostérie.** (haut) Chaque site peut être inactif ou actif. En combinant trois sites, la protéine a donc huit états différents. (bas) La taille des représentations des structures est corrélée à leur probabilité de formation (les grandes conformations sont très susceptibles de se formées). (bas, à gauche) L'effecteur 1 diminue l'énergie de conformations où le site fonctionnel est actif (flèches pointillées). Il augmente ainsi la probabilité du site fonctionnel à être actif : la conformation de plus basse énergie ayant le site fonctionnel actif grâce à la présence de l'effecteur1. L'effecteur 1 est dans ce cas activateur. (bas, à droite) L'effecteur 2 est préalablement fixé. L'ajout de l'effecteur 1 va diminuer l'énergie des mêmes conformations que dans la figure de gauche (flèches pointillées). Cependant, une seule des deux conformations de plus basse énergie présente le site fonctionnel actif alors que c'est le cas des deux conformations de plus basse énergie en présence de l'effecteur 2 uniquement. La probabilité de présence d'un site fonctionnel actif est donc diminuée : l'effecteur 1 passe d'un rôle activateur à un rôle répresseur. Figure reproduite avec la permission des auteurs[509].

Comparés à un modèle structural, les effets de modulation allostérique complexes décrits ci-dessus ne peuvent adhérer à ces approches car elles ne permettent pas de décrire les comportements dynamiques observés. À l'inverse, une approche statistique basée sur un modèle à ensemble permet ce degré de complexité où deux sites peuvent être couplés positivement lorsqu'un sous-ensemble domine, ou négativement lorsqu'un autre sous-ensemble domine. Dans ce contexte, le remodelage du paysage énergétique lié à l'introduction d'une nouvelle perturbation va modifier le sous-ensemble qui domine, et potentiellement transformé un effet '*activateur*' en effet '*répresseur*', ou inversement. Ces perturbations ont de nombreuses origines : modifications post-traductionnelles (SUMOylation, phosphorylation, acetylation), épissage alternatif et production de formes tronquées, liaison d'un second effecteur allostérique, *etc*.

La régulation allostérique liée à des mutations est particulièrement intéressantes car les mutations peuvent remodeler l'activité d'une protéine. L'étude de ces mutations constitue ainsi un champ important dans la recherche clinique et pharmacologique. Clarkson *et al*. ont observés deux types de propagation allostérique en réponse à des mutations ponctuelles dans la sérine protéase Eglin c[513]. La réponse se déplace soit sous la forme d'un chemin contigu de changements dynamiques, soit sous la forme de changements dispersés associés à des changements subtils de la chaîne principale de la protéine. Schrank *et al.* ont ensuite conçus des protéines mutées qui affectent des sites fonctionnels distants en utilisant ces concepts[514]. Les simulations de dynamique moléculaire des formes sauvages et mutées des protéines Abl et du récepteur au facteur de croissance épithélial (*Epithelial Growth Factor*, *EGF*) ont également mis en évidence les changements du réseau allostérique induits par des mutations oncogéniques[515]. Dans ces cas, la communication allostérique est décrite en termes de couplage dynamique entre des éléments rigides et capable d'adapter différentes conformations. L'hypothèse formulée est que ces éléments structuraux forment un réseau dynamique d'interactions fonctionnelles qui contrôlent la communication longue portée et l'activation allostérique de ces protéines kinases.

Dans les premiers modèles allostériques MWC et KNF, la transmission d'une information d'un site à un autre était décrite par **un chemin de communication** (une succession de résidus interagissant les uns avec les autres) unique et bien défini, et entraine un changement conformationnel sur le site de liaison fonctionnel. Les résidus qui constituent ce chemin sont dès lors considérés comme des résidus allostériques. Le signal est alors transmis par ces résidus allostériques en établissant une succession de liaisons non-covalentes, permettant la transmission d'une information entre les sites, de proche en proche. Cependant, de nombreux phénomènes biologiques ne peuvent être expliqués par cette approche. En particulier, on ne peut expliquer l'apparition des effets induits par une mutation lorsque le résidu muté ne se situe ni dans le site fonctionnel ni appartient aux résidus allostériques. Le modèle à ensemble a là aussi apporté un nouveau concept permettant d'expliquer ces phénomènes qui postule l'existence des plusieurs chemins, dont l'importance est variable, qui permettent de faire la jonction entre les différents sites. La transmission de l'information suite à la liaison (covalente ou non-covalente) d'un effecteur ou à la présence d'une mutation va alors transiter le long de

plusieurs chemins. À la manière d'une onde de choc, une perturbation induit un « stress » qui va se propager en empruntant des voies privilégiées, dépendantes des conditions. L'impact des mutations des résidus qui constituent ces chemins de communications vont ainsi permettre de discriminer les voies de communications importantes et secondaires. Si la mutation engendre un effet majeur sur l'activité, alors elle fait partie d'une voie de communication importante, et inversement, des effets mineurs sont liés à des voies secondaires.

Cependant, les résidus participant directement aux chemins de communication ne sont pas les seuls à pouvoir être impliqués dans la communication allostérique. Les effets des différents perturbateurs (mutation, ligand, …) étant liés au sous-ensemble conformationnel dominant, un résidu pouvant modifier l'équilibre entre sous-ensembles va impacter la régulation de la protéine en privilégiant certaines transitions allostériques. Certaines mutations vont ainsi piéger un système dans une zone énergétiquement stable, et empêcher le déplacement de l'équilibre.

La caractérisation de ces chemins de communication peut se faire sous forme d'un réseau d'interactions entre les différentes régions du système, et les voies de communication vont se croiser au niveau de résidus clé avant de se propager dans les différentes régions, présentant un comportement pseudo-rigide[516] et connectés entre eux. Les mouvements de la protéine sont reliés à l'organisation de ce réseau de communications dont la perturbation influera réciproquement sur la dynamique du système et induira les changements conformationnels au niveau du site fonctionnel.

La notion de chemin de communication implique un transfert d'information à la fois dans le temps et dans l'espace. Dans l'espace, car les chemins consistent en une succession de contacts dynamiques entre résidus qui transfèrent une charge énergétique générée par l'effecteur/perturbateur. Les chemins de communication allostérique peuvent être vus comme une succession d'interactions inter-résidus qui, mis bout à bout, relient le site de liaison de l'effecteur au site actif ou un autre site. Une difficulté de cette conception réside dans l'évaluation de la contribution de chaque interaction entre résidus. Une approche statistique (de manière similaire au modèle à ensemble des différents états de la protéine) pourrait être une voie d'évaluation de la fréquence de passage d'un chemin allostérique à travers chaque résidu. Les résidus allostériques clés sont alors ceux présentant la plus haute fréquence, ou une fréquence supérieure à un seuil.

Dans les faits, *in vivo*, la présence d'un seul effecteur allostérique et un seul site fonctionnel (la situation la plus simple) est aussi la plus rare. La dynamique cellulaire est telle que chaque protéine est soumise à des interactions simultanées de la part de multiples intervenants pouvant se fixer sur différents sites effecteurs. L'expression de « complexes macromoléculaires » est ainsi devenue une expression de plus en plus employée afin de rendre compte de la complexité de l'environnement cellulaire. De multiples évènements conjoints conduisent à générer de multiples signaux allostériques sur une seule et même protéine. Les effets allostériques observés, qu'ils se reflètent sur la dynamique ou la structure de la protéine,

dépendent donc de la nature de la perturbation ET du site où il vient se fixer. Dans ce cadre, le moindre changement d'environnement peut agir comme un effecteur : une phosphorylation, la liaison covalente d'une molécule (SUMO, ubiquitine), la fixation non-covalente d'une molécule (inhibiteur, acide nucléique), *etc*.

Les protéines impliquées dans les voies de signalisation cellulaires sont particulièrement sujettes à ce type d'activation, et constituent un bon exemple de combinaison de signaux multiples. Ces systèmes sont souvent la cible de modifications post-traductionnelles, d'interactions au sein de complexes macromoléculaires et/ou de changement de compartiment cellulaire. Un exemple de protéine qui subit ces modifications de manière simultanée, et liée à la régulation de l'activité des protéines STAT, est la famille des SOCS (*Suppressor of cytokine signaling*)[517]. La synthèse des SOCS est régulée par les cytokines, dont l'activité est ainsi modulée par les SOCS dans une boucle de régulation négative. L'activité des SOCS est par ailleurs stimulée par d'autres signaux, ce qui suggère de multiples interfaces capables d'interagir avec de multiples partenaires : les protéines SOCS, qui sont un élément de la machinerie moléculaire ligase E3, peuvent se lier aux élongines B et C par un site tandis qu'elles se lient à différents substrats par un autre site, et ainsi subir des modifications post-traductionnelles. Ainsi, les filets d'énergie induits par chaque perturbation vont se propager à la manière d'une onde, et fusionner comme des vagues à la surface de l'eau. Par ces interactions, elles vont altérer les effets des unes et des autres, et entraîner des modifications spécifiques au niveau du site fonctionnel. Ces chemins restent inaccessibles à l'observation directe et par conséquent, ces modifications spécifiques sont aujourd'hui impossibles à suivre par des méthodes expérimentales. Cependant, l'observation indirecte des effets allostériques permettent de tirer des conclusions relatives à ces chemins.

Le modèle allostérique actuel explique la complexité des phénomènes allostériques. Une protéine est un système multidimensionnel qui comprend plusieurs états stables et métastables. La transition d'un état à un autre est décrite par la surface d'énergie sous-jacente, qui peut être modifiée par la présence d'un élément nouveau. Par ailleurs, le modèle introduit une hiérarchisation de la protéine avec l'introduction des différents sites – sites fonctionnels qui portent les fonctions régulées, et sites de liaison des effecteurs qui modulent l'activité des sites fonctionnels. Avec l'intégration de ces notions, le modèle offre la possibilité d'observer des effets allostériques qui divergent suite à la stimulation d'une même perturbation, en fonction du sous-ensemble conformationel qui domine dans le paysage énergétique et de l'état de chacun des sites (*cf.* Figure 21), chaque liaison ou modification covalente (modification post-traductionnelle, liaison à une molécule d'ubiquitine, *etc.*) pouvant constituer un élément de perturbation – un effecteur. Ces perturbations altèrent le réseau interne d'interactions de la protéine, et modifient les **voies de communications**. Cette modulation, qui peut par ailleurs être positive (couplage positif de deux sites) ou inhibitrice (couplage négatif de deux sites) et qui est susceptible aux effets des autres perturbations, constitue l'origine de la modulation dynamique des protéines allostériques.

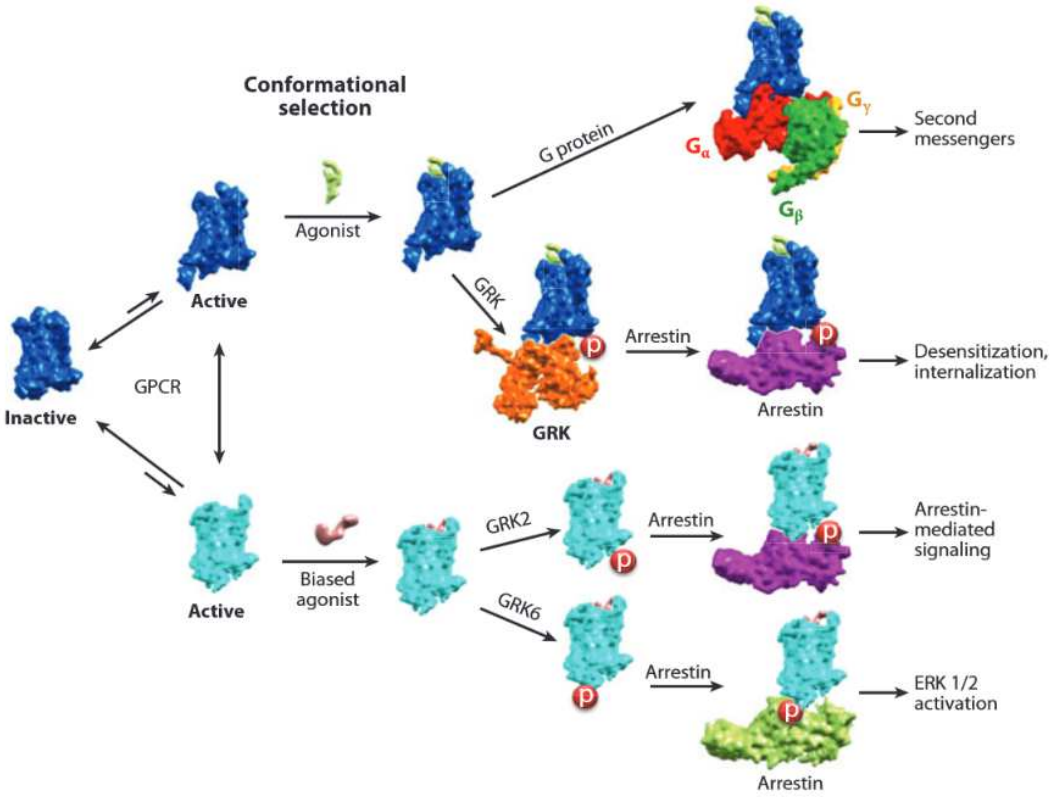


Figure 21: **La diversité des voies de signalisation est accentuée par les effets allostériques dans les RCPGs.** En fonction du sous-ensemble qui domine le paysage conformationnel, la fixation d'un ligand donné va favoriser une conformation du récepteur, et permettre ainsi l'activation de voies de signalisation différentes (en bleu ou en cyan). Figure reproduite avec la permission des auteurs[518].

Les protéines sont par essence même des systèmes allostériques. Le cytoplasme présente une très forte concentration de protéines, conduisant inévitablement à des interactions entre protéines. Les perturbations liées aux interactions entre les protéines restent cependant faibles, car très peu spécifiques. La fixation d'un effecteur allostérique va en effet impacter d'une manière plus radicale le réseau allostérique d'une protéine et déplacer l'équilibre conformationnel d'un sous-ensemble vers un autre. L'étude détaillée de ces différents niveaux d'expression de l'allostérie peut ainsi dévoiler de nouvelles stratégies de modulation de l'activité, avec un gain en sélectivité potentiellement importante.

C. Application de l'allostérie dans la recherche de composés actifs

Les modulateurs allostériques (cytokines, neurotransmetteurs, *etc.*) induisent et améliorent la reconnaissance, l'amplification et la transmission d'un signal, sans entrer en compétition avec les ligands endogènes. Ainsi, les modulateurs allostériques adoptent une approche complètement différente des effecteurs orthostériques (qui entrent en compétition avec les ligands endogènes sur un site d'une cible donnée), en offrant potentiellement la possibilité de faire varier finement le degré d'activation[519] (ou d'inactivation) d'une protéine. Ils sont donc une alternative plus fine pour la modulation de l'activité des systèmes biologiques, qui peut passer inaperçu en l'absence de ligands. En effet, en perturbant de manière moindre la

régulation de la protéine qu'ils ciblent, les modulateurs allostériques préservent d'autant la régulation des processus cellulaires, notamment dans le cas de protéines multifonctions, en augmentant ou diminuant la force des voies de communication.

La famille des récepteurs couplés aux protéines G (*RCPGs*) est la principale cible des composés actifs commercialisés à l'heure actuelle. Les membres de cette famille sont activés par la fixation dont ligand qui assure, à partir d'un état latent, l'activation du récepteur, donc de la transmission d'un signal. Les inhibiteurs orthostériques ont accès au site de fixation des ligands et l'obstruent, alors que les ligands allostériques se fixent sur un second site. Le site allostérique du récepteur muscarinique M2, distinct du site orthostérique, reconnait la gallamine. Ce site reconnaît également les ligands du site orthostérique lorsqu'ils sont en excès[520] et suggère une coïncidence de certains chemins de propagation. La survenue de phénomènes allostériques au sein des GPCRs a été montré par leur capacité à se dimériser sous l'action de ligands ciblant le site orthostérique. Ainsi, tous les sites pourraient être allostériques, et le résultat final de la liaison d'un ligand serait le décalage de l'équilibre thermodynamique de l'ensemble vers un autre ensemble pertinent d'un point de vue pharmacologique *via* la stabilisation d'un état conformationnel donné. Le LPI805 est un inhibiteur non-compétitif de la liaison d'un agoniste orthostérique, la neurokinine A (NKA), aux récepteurs de la tachykinine NK2 (*NK2Rs*), la liaison NKA-NK2R favorisant la conformation A2L du récepteur[521]. La présence de LPI805 déplace l'équilibre conformationnel vers la conformation A1L, et inhibe en conséquence la réponse du récepteur à l'AMP cyclique, alors que la réponse au calcium est augmentée[521]. La régulation allostérique mise en place est donc sélective, les fonctions n'étant pas impactées de la même manière. Enfin, une étude récente a montré la régulation positive et sélective des récepteurs nicotinique à l'acétylcholine (*nAChRs*) α7 par le PNU-120596[522]. Ce produit augmente et prolonge l'activation des récepteurs α7, mais n'induit pas leur activation en l'absence d'agoniste[522]. Ainsi le ciblage de sites allostériques présentes plusieurs avantages potentiels : en plus de proposer de nouveaux sites d'interactions, la modulation du signal de transduction est potentiellement plus fine et sélective, certaines fonctions étant positivement régulées quand d'autres s'en trouvent négativement régulées. Cependant, des difficultés restent à surmonter. Le principal obstacle est la diversité des effets en fonction de la conformation dominante.

À l'instar des inhibiteurs pharmacologiques, les modulateurs allostériques peuvent être covalents (lié chimiquement avec le résidu d'un site de liaison de la cible) ou non covalents (fixé par des interactions non-covalentes). Les interactions liés à un type de fixation ou à un autre vont entrainer *a priori* des effets plutôt de type irréversible (liaison covalente) ou réversible (liaison non-covalente). Cependant, la distinction entre ces deux classes de modulateurs n'est pas aussi nette qu'elle n'y parait au premier abord. La liaison covalente d'un modulateur de ce type peut en effet être rompue, alors que le changement conformationnel induit par un ligand non-covalent (entre un état actif et un état inactif d'une protéine, bien que réversible), peut être très difficile à annuler. Si la surface énergétique entre la conformation inactive et la conformation active présente une barrière importante, il est en effet peu probable que la protéine retourne à l'état actif rapidement : l'inhibition obtenue est durable. Une inhibition

irréversible reste néanmoins plus facile à obtenir à l'aide d'un ligand covalent. Ainsi, les modulateurs allostériques partagent les mêmes mécanismes d'action que les mutations. Une perturbation nouvelle (mutation ponctuelle ou insertion/délétion) modifie le réseau d'interactions internes à la protéine, et entraine la propagation d'un nouveau signal cellulaire[523]. Ce signal, qui peut être délétère ou bénéfique, présente une hétérogénéité liée aux multiples sous-ensembles conformationnels de la protéine. La prédiction des effets d'une perturbation donnée sur l'activité d'une protéine constitue donc un défi majeur en biologie.

Chapitre 2 : Méthodes & méthodologie

Les macromolécules telles que les protéines et les acides nucléiques jouent un rôle central dans les processus biologiques, physiologiques ou pathologiques. Elles constituent un domaine de recherche très vaste qui s'appuie très largement sur des données expérimentales afin de décrire ces processus, ou de générer de nouvelles hypothèses. Dans ce cadre, la biologie structurale permet d'étudier les systèmes biologiques à l'échelle atomique, et ainsi de compléter la description de phénomènes macroscopiques à une échelle de dimension inférieure. Cependant, des données structurales ne sont pas disponibles pour l'ensemble des systèmes étudiés, et ne fournissent que peu souvent des informations de type dynamique, pourtant essentielles dans les processus biologiques. Ainsi, la biologie computationnelle s'est développé afin d'apporter de nouvelles approches permettant de combler en partie ce manque. Dans ce chapitre, les méthodes, programmes et paramètres utilisés seront explicités, et replacés dans le contexte plus général décrit dans le chapitre 1.

I. Modélisation par homologie

Explorations des bases des données : Les séquences primaires des formes humaines des protéines *Signal Transducer and Activator of Transcription* STAT5a et STAT5b ont été obtenues à partir de la base de données des protéines du NCBI (*National Center for Biotechnology Information*, http://www.ncbi.nlm.nih.gov/protein). Une recherche a permis d'identifier les séquences canoniques humaines de ces deux protéines sous les références *NP_003143.2* et *NP_036580.2* pour STAT5a et STAT5b respectivement. Ces deux séquences présentent une identité de séquence de 92,88% (*cf.* Tableau 1). À partir de ces données, une

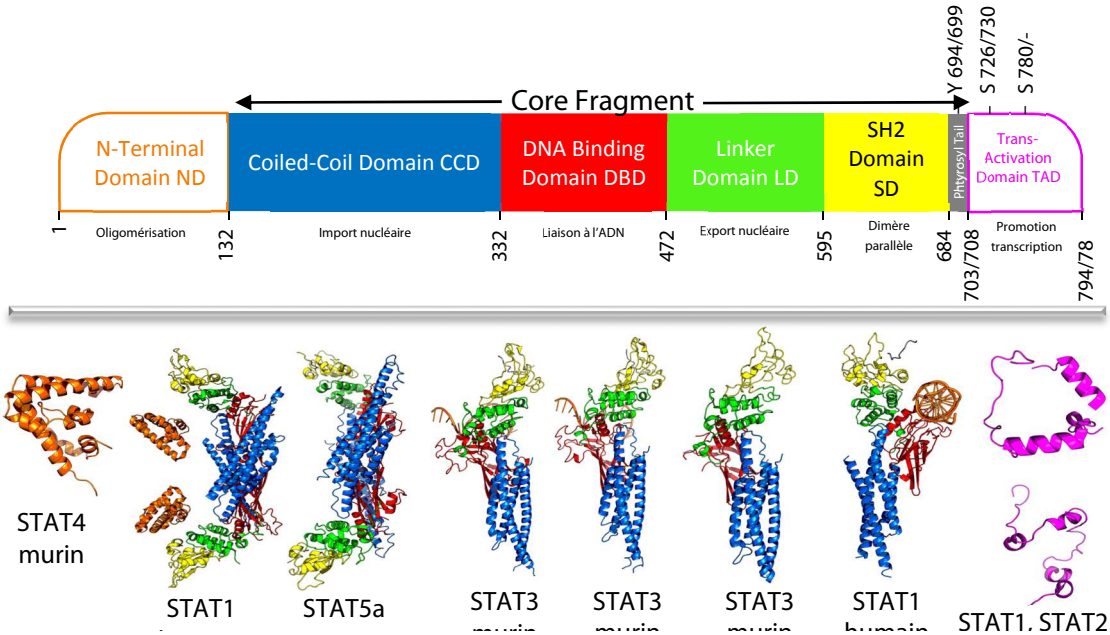


Figure 22 : **Structures cristallographiques disponibles pour les protéines STATs, colorées par domaine.** (haut) Représentation schématique des domaines des STATs, les principales fonctions sont indiquées. Les numérotations des résidus sont celles de STAT5. Lorsque la numérotation de STAT5a et STAT5b diffèrent, elles ont le format STAT5a/STAT5b. (bas) Représentation tridimensionnelle de quelques-unes des structures PDB, colorées par domaine. Le schéma des couleurs est conservé par rapport au haut.

recherche sur la séquence protéique complète a été effectuée sur l'outil *protein blast* (*protein Basic Local Alignment Search Tool,* http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome) en comparant les séquences primaires au séquences des structures protéiques référencées dans la *Protein Data Bank* (PDB, *cf.* II.B.2 du chapitre 1). Cette recherche a permis d'identifier (i) aucune structure répertoriée dans la PDB n'est celle de STAT5a ou STAT5b dans sa forme complète ; (ii) aucune structure modèle correspondant à la séquence primaire complète d'une des protéines STATs ; (iii) 10 structures cristallographiques ou RMN présentant une bonne correspondance, bien que partielle, aux séquences primaires complètes. Ces structures et leurs principales propriétés sont résumées dans le Tableau 3. L'analyse des structures identifiées nous a permis d'identifier les domaines des STATs pour lesquels nous disposons du plus grand nombre d'informations structurales (*cf.* Figure 22).

Les domaines N- et C-terminaux des protéines STATs sont peu caractérisés. Le domaine N-terminal est caractérisé dans seulement deux structures (*cf.* Tableau 3), soit sous la forme d'un dimère du domaine isolé de STAT4 (1BGF[138]), soit dans un cristal d'une forme tétramérique anti-parallèle de STAT1 tronquée au niveau du domaine TAD (1YVL[117]). Aucune donnée n'est donc disponible concernant le positionnement du domaine N-terminal dans les formes monomériques ou les dimères parallèles. De plus, le domaine N-terminal est relié au CCD par une longue boucle non structurée, ce qui rend particulièrement difficile la prédiction du positionnement relatif du domaine ND par rapport au reste de la protéine. Enfin, le rôle du domaine ND est essentiellement un rôle d'oligomérisation, que ce soit pour la formation de tétramères liés à l'ADN ou dans la transition entre les dimères anti-parallèles et parallèles [117,524]. Dans ce cadre, nous n'avons pas inclus ce domaine pour la modélisation de STAT5. Le domaine C-terminal n'est que partiellement résolu, et toujours lorsqu'il est lié à d'autres partenaires protéiques [525,526]. La prédiction des structures secondaires à partir de la séquence primaire pour ce domaine ainsi que pour la boucle qui porte le résidu phosphotyrosyl a mis en avant l'absence de structures secondaires sur ces régions de la protéine, à l'exception notable d'une hélice α dans le domaine TAD. La fonction principale de ce domaine est d'assurer le recrutement de partenaires protéiques afin de permettre la formation d'un complexe de transcription. Le domaine TAD n'est par ailleurs pas requis pour la formation des dimères de STATs, certaines formes tronquées participant par ailleurs à la régulation de l'activité de STAT5[527]. Nous avons donc exclu le TAD de nos modèles.

Une seconde exploration a été réalisée par *protein BLAST* sur les séquences primaires de STAT5 tronquées (sans ND et TAD), correspondant aux résidus 136 à 703 pour STAT5a et 136 à 708 pour STAT5b. Aucune nouvelle structure couvrant l'ensemble du *CF* ou d'un des domaines n'a été détectée au cours de cette seconde analyse. Deux structures ont été choisies comme supports (*template*) pour générer les modèles de STAT5a et STAT5b humain par homologie:

- 1Y1U[118], résolue à partir d'un cristal de STAT5a murin non-phosphorylé. La structure est un dimère anti-parallèle, dont les monomères interagissent par les domaines CCD et DBD. Cette

structure présente une excellente identité de séquence avec STAT5a et STAT5b humains : 97,6 et 95,9% respectivement. Les domaines ND et TAD sont absents de la protéine, alors que les résidus 129 à 137 (boucle reliant les domaines ND et CCD), 424 à 432 (boucle du domaine DBD) et 690 à 712 (queue phosphotyrosyl et début du domaine TAD) n'ont pas pu être localisés. La résolution de cette structure est de 3,2 Å, une résolution faible.

- 1BG1[132] comprend les résidus 127 à 722 des deux monomères du dimère parallèle de STAT3β murin lié à l'ADN, tronqué au niveau du domaine C-terminal. Il s'agit de la structure du *Core Fragment* résolue à la meilleure résolution (2,3 Å). Les résidus 127 à 135 (boucle N-terminale), 184 à 194 (boucle reliant les hélices α1 et α2), 688 à 701 (boucle constituée de l'extrémité C-terminale du domaine SH2 et de l'extrémité N-terminale de la queue du résidu phosphotyrosyl) et 717-722 (domaine TAD) n'ont pas été déterminés. La séquence de cette structure présente une identité de séquence de 30% avec les deux STAT5.

Tableau 3: Structures de protéines STAT répertoriées dans la PDB (www.rcsb.org, [441]).

Code PDB	Isoforme	Résidus résolus	Résolution (Å)	ADN	Phosphorylation de la tyrosine	Référence
1BF5¤	STAT1 Humain	L136-H182 ; L197-R683 ; G700-S710	2.9	Double-brin	Oui	[133]
1YVL¤	STAT1 Humain	S2-Q126 ; T133-E181 ; A188-N414 ; I425-K544 ; K550-Q621 ; E625-R683	3.0	Non	Non	[117]
2KA6§	STAT1 Humain	G706-V750	NA	Non	NA	[526]
2KA4§	STAT2 Humain	G782-S838	NA	Non	NA	[526]
1BG1¤	STAT3 Murin	V136-D184 ; S194-R698 ; A702-F716	2.3	Double-brin	Oui	[132]
3CWG¤	STAT3 Murin	V136-K180 ; M200-K370 ; R379-C418 ; I431-W623 ; K631-N646 ; F650-R688	3.1	Non	Non	[121]
4E68¤	STAT3 Murin	V136-D184 ; S194-R688 ; A702-F716	2.6	Double-brin	NA	[134]
1BGF¤	STAT4 Murin	G0*-I123	1.5	Non	NA	[138]
1Y1U¤	STAT5 Murin	S138-R423 ; E433-A690	3.2	Non	Non	[118]
1OJ5¤	STAT6 Humain	L795-E808	2.2	Non	NA	[525]

La marque * indique la présence d'un motif Glycine-glycine à la place du résidu 1 (1BGF), d'où une numérotation à partir de 0. Le sigle ¤ marque les structures cristallographiques alors que le sigle § marque les structures obtenues par RMN. NA = Non Applicable.

Choix de la séquence d'ADN à modéliser: Le fragment d'ADN reconnu par les protéines STATs varie bien évidemment d'une STAT à une autre, mais un motif palindromique commun de type $TTCN_{2-4}GAA$ est généralement admis comme étant la séquence consensus des facteurs de transcription STAT[11,528]. STAT3 et STAT5 présentent la même préférence pour des séquences en N_3, qui sont des séquences où les fragments de motif TTC et GAA sont séparés par trois nucléotides [524,529].Un motif de ce type est donc logiquement présent dans la structure 1BG1 : TTCCGGGAA. Le motif central composé de trois nucléotides (CGG) est dit de « faible » affinité, puisqu'il n'apporte pas de spécificité de reconnaissance entre les différentes STATs. Cependant, il a été montré que des différences significatives existent entre les séquences reconnues au niveau de ce motif central. Ehret et collaborateurs ont ainsi établi une carte des séquences préférentiellement reconnues par STAT1, STAT5a, STAT5b et STAT6[530]. STAT5a et STAT5b partagent la même séquence préférentielle – TTCTTAGAA. Modeller ne pouvant prendre en charge la modélisation de résidus nucléotidiques, la mutation des bases azotées CGG → TAA a été réalisée à l'aide de l'outil *Coot*[531].

Modélisation par homologie : Les modèles de STAT5 (monomères et homodimères) ont été générés par homologie, par l'utilitaire Modeller 9v10[532], à partir de l'alignement des séquences des deux supports structuraux (1BG1 et 1Y1U) avec les deux séquences cibles (STAT5a et STAT5b), présenté dans la Figure 23. La modélisation de la boucle entre la (phospho-)tyrosine et le domaine SH2 est difficile car cette zone est très divergente dans les structures supports. Pour limiter la nature de ce problème, nous avons choisi dans le cas du dimère de ne pas appliquer de contraintes de symétrie en sus des contraintes spatiales générées automatiquement par Modeller. Les modèles de STAT5 générés ont été évalués à l'aide de la fonction de score DOPE (*Discrete Optimized Protein Enegy*)[533], en gardant à l'esprit que cette fonction ne peut différencier les systèmes phosphorylés dessystèmes non-phosphorylés.

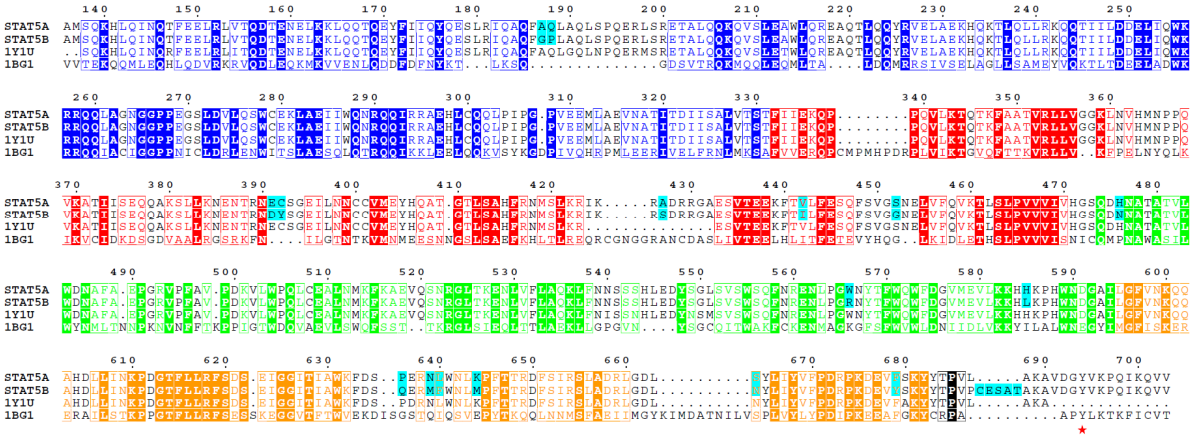


Figure 23: **Alignement de séquence utilisé pour l'étape de modélisation par homologie**. Les couleurs permettent de distinguer les différents domaines (bleu = CCD, rouge = DBD, vert = LD, jaune = SH2D et noir = queue phosphotyrosyl). Un fond uni indique une identité des séquences, l'encadré une similarité. Le fond cyan montre les points de mutation entre les séquences de STAT5a et STAT5b. Le résidu de tyrosine critique est indiqué par l'étoile rouge.

Au total, 100 modèles ont été générés, pour quatre espèces monomériques (STAT5a non-phosphorylé et phosphorylé, STAT5b non-phosphorylé et phosphorylé) et deux espèces dimériques (STAT5a phosphorylé et STAT5b phosphorylé) liées à l'ADN. Pour chaque espèce, le modèle présentant le plus bas score DOPE a été inspecté manuellement en utilisant des outils graphiques (VMD[446] ou PyMOL[442], et retenu comme ***modèle de départ*** s'il ne présentait pas d'artefacts structuraux au niveau C-terminal (en particulier, un « nœud » dans une boucle dans les formes monomériques, boucle d'une sous-unité qui traverse le domaine SH2 de l'autre sous-unité dans les formes dimériques). Les modèles ont ensuite été contrôlés par Procheck[534], afin de vérifier que les angles dièdres de la chaîne principale sont correctement modélisés. Pour nos modèles, moins de 2% des résidus sont situés dans la zone la moins favorable du graphique de Ramachandran.

L'ajout des atomes d'hydrogène et l'évaluation de l'état de protonation des résidus à un pH neutre de 6,8 a été réalisé par le serveur web H++ v3.0 (http://biophysics.cs.vt.edu/H++)[535–537]. L'analyse des structures cristallographiques disponibles de dimères des protéines STATs liés à l'ADN [132–134] et les résultats des études de dynamique moléculaire portant sur STAT3[538] nous ont permis d'identifier le résidu N460 chez STAT1 et son équivalent N466 chez STAT3 comme primordiaux pour la reconnaissance spécifique du motif $TTCN_3GAA$. Dans les protéines STAT5, le résidu équivalent aux résidus N460 de STAT1 et N466 de STAT3 est le résidu H471. Pour la modélisation de STAT5, trois états de protonation distincts de ce résidu H471 ont été évalués. Sans indication concernant l'état de la protonation de ce résidu, et étant donné son importance biologique, nous avons choisi de simuler les trois états de l'histidine, protonée en δ, en ε, ou sur les deux sites (*cf.* Figure 24).

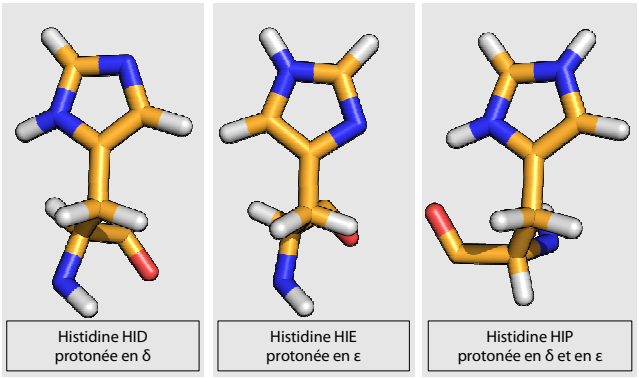


Figure 24 : **Les différents états de protonation du résidu d'histidine.**

II. Minimisations des modèles et simulations de dynamique moléculaire

A. Champ de forces

Les simulations de dynamique moléculaire (*cf.* II.D.2 du chapitre 1) ont été réalisées avec la version 4.5 de GROMACS[539], en utilisant le champ de forces amber ff99SB*-ILDN[540–542], sauf indication contraire. Ce champ de forces regroupe trois extensions successives dans le but d'améliorer le champ de forces ff99. Le premier travail de Hornak et collaborateurs[540] a visé à améliorer les paramètres liés aux angles dièdres φ et ψ des résidus, produisant le champ de forces ff99SB. Par la suite, une correction énergétique pour les chaînes principales a été introduite afin de corriger la balance entre les différents types de structure secondaire[541], générant le champ de forces obtenu ff99SB*.

Nous avons réalisé initialement quelques essais préparatoires de simulations de dynamique moléculaire de STAT5 sans utiliser ce terme correctif, ce qui a donné des résultats très délétères sur la stabilité des longues hélices α. Les hélices du CCD de STAT5 notamment se brisaient en effet rapidement en leur centre, et nous ne pouvions pas conserver la courbure de l'hélice α2, pourtant bien caractérisée dans la structure cristallographique 1Y1U[118]. Enfin, le travail de Lindorff-Larsen et collaborateurs[542] a porté sur l'amélioration des chaînes latérales des résidus d'isoleucine (I), leucine (L), aspartate (D) et asparagine (N) a procuré le champ de forces dénommé ff99SB*-ILDN.

Les champs de forces sont des outils indispensables pour réaliser des simulations de dynamique moléculaire. Ils établissent le lien entre la description de la nature des atomes du système et les théories physico-chimiques qui régissent les interactions atomiques donc leur déplacement. Ainsi, chaque atome du système se voit attribuer un type atomique associé à un jeu de paramètres qui permet le calcul des forces. Un ensemble typique de paramètres comprend des valeurs pour la masse atomique, le rayon de van der Waals et la charge partielle de chaque type d'atomes, et les valeurs d'équilibre des longueurs de liaison, des mesures d'angles plans et dièdres pour des paires, triplets et quadruplets d'atomes liés, et les valeurs de constante de force pour chaque potentiel du champ de forces. De la justesse du type atomique associé à un atome donné dépend la qualité de la simulation atomique. Pour des simulations de macromolécules, l'association atome/type atomique se fait de manière automatique en se basant sur la structure des protéines et/ou acides nucléiques. Les acides aminés et les nucléotides sont des éléments répétitifs et de composition constante, ce qui limite, au regard de la diversité du vivant, le nombre de types d'atomes nécessaire pour décrire un système biologique. Cependant, des éléments non standards, comme un ligand ou un groupement phosphate, peuvent être nécessaires à la description d'un complexe macromoléculaire. Un nouveau jeu de type atomique (donc de paramètres associés) doit souvent être ajouté au champ

de forces, les types atomiques intégrés de base au champ de forces ne permettant pas de décrire les propriétés du ligand. Dans le cadre de notre projet, le résidu de tyrosine phosphorylé nécessite l'ajout des paramètres associés. Nous avons introduit manuellement ces paramètres, à partir des données issues des travaux de Homeyer et collaborateurs[543], dans le dossier du champ de forces ff99SB*-ILDN. Les paramètres des champs de forces ont ainsi deux origines : ils sont extraits des bases de données (mesures expérimentales, empiriques), et incluent parfois des paramètres issus de calculs théoriques basés sur les théories quantiques. Les termes de champ de forces empiriques et semi-empiriques sont ainsi employés.

B. Minimisations des modèles de STAT5 dans le vide

La modélisation par homologie génère des modèles tridimensionnels de STAT5 *via* l'utilisation de contraintes spatiales visant à ne pas produire de clashs stériques. Cependant, dans des systèmes présentant un nombre de degré de liberté aussi élevé qu'une structure protéique, l'étape de modélisation par homologie peut produire des structures présentant des artefacts de type interpénétrations atomiques. De plus, et en dépit d'une bonne identité de séquence entre les structures support (1Y1U et 1BG1) et STAT5a/b, ces protéines pourraient ne pas partager le même minimum énergétique. La minimisation des structures générées va ainsi permettre de produire des structures stables et dépourvues de gènes stériques. Cette première étape de modélisation a été réalisée dans le vide, et a consisté en deux minimisations successives des modèles utilisant deux algorithmes de minimisation énergétiques courants:

- l'algorithme de plus grande pente (*steepest descent*) prend en entrée les coordonnées atomiques R^n d'un système et son énergie E^n. Chaque atome va être déplacé le long d'un vecteur qui suit la direction de plus grande pente de la surface d'énergie D^n, produisant un nouveau jeu de coordonnées :
  $R^{n+1} = R^n + \frac{F^n}{max(|F^n|)} D^n$, dont l'énergie est E^{n+1}. Si $E^{n+1} < E^n$, la position $n+1$ est considérée comme plus stable et les nouvelles coordonnées sont acceptées et $D^{n+1} = 1{,}2D^n$. Sinon, les coordonnées R^nsont conservées et $D^{n+1} = 0{,}2D^n$. Cet algorithme converge rapidement vers un minimum énergétique proche, dans lequel il va osciller. Pour sortir de cette boucle, on s'appuie sur deux critères de convergence qui, si l'un est atteint, va finir l'étape de minimisation. Ces critères de convergence sont d'une part le nombre d'itération, fixé à 50 000, et le gradient d'énergie minimum entre deux positions successives, fixé à 100 kJ/mol/nm.
- le second algorithme utilise la méthode du gradient conjugué, qui est plus efficient à proximité d'un minimum local du modèle. Cette approche utilise à la fois le gradient d'énergie et la variation du gradient entre les deux dernières positions pour déterminer les nouvelles positions du système. Les critères de convergence sont les mêmes, mais ont été fixés à 50 000 pour le nombre de pas d'itération maximum, et à 10 kJ/mol/nm afin d'aboutir à une structure minimisée correctement.

C. Modèle des molécules d'eau et d'ions

Le milieu intracellulaire est très fortement peuplé (protéines, ligands,…), mais le cytoplasme reste avant tout un milieu essentiellement aqueux. Les simulations de dynamique moléculaire des protéines et macromolécules biologiques se doivent de mimer cet environnement afin de reproduire le plus fidèlement possible le milieu cytosolique. Si l'ajout d'un milieu implicite (possédant une constante diélectrique fixe) est possible, l'ajout de molécules d'eau explicites est préférée car elle rend mieux compte des interactions protéines – solvant et donc des effets dynamiques du système. Cependant, cette précision se fait au dépend du volume de calcul, qui est bien plus important dans le cas des solvants explicites.

Plusieurs modèles de molécule d'eau existent, qui diffèrent par leurs propriétés physico-chimiques, plus particulièrement par la longueur des liaisons entre les atomes d'hydrogène et d'oxygène (H—O), l'angle entre les atomes d'hydrogène, d'oxygène et d'hydrogène (H—O—H) , ainsi que par les charges partielles et électrostatiques des atomes. Les modèles à trois points (TIP3P, SCP/E[544,545], *etc.*) font coïncider la charge ponctuelle de l'atome d'oxygène avec le centre de l'atome, alors que les modèles à quatre points (TIP4P et BF[544,546], par exemple) ou à cinq points (TIP5P[547]) placent les charges en dehors des atomes afin de mieux représenter la charge électrostatique et les doublets non-liants de l'atome d'oxygène. Si les derniers modèles sont plus précis, ils sont aussi plus couteux en temps de calcul. Nous avons utilisé dans les simulations le modèle TIP3P, et avons ajouté des ions sodium (Na^{2+}) ou chlorure (Cl^{-}) afin que la charge totale du système à simuler (protéine + acides nucléiques + eau) soit nulle. Enfin, les conditions périodiques aux limites du système ont été appliquées dans les trois dimensions afin que la boîte centrale soit entourée de ses propres images (*cf.* Figure 25). Si un atome sort des limites de la boîte au cours de la simulation, son image rentre par le côté opposé de la boîte centrale, ce qui assure un nombre constant d'atomes dans la boîte.

La taille des systèmes simulés, comprenant les atomes d'eau, les ions, les protéines et l'ADN (dans le cas des systèmes dimériques de STAT5 liés à l'ADN), est de 200 000 atomes environ pour les formes monomériques, et de 320 000 atomes pour les dimères liés à l'ADN. Les monomères de STAT5a comptent 568 résidus protéiques, alors que les fragments de STAT5b comptent 573 résidus. Les doubles-brins d'ADN simulés sont composés de 18 nucléotides, formant 17 paires de bases.

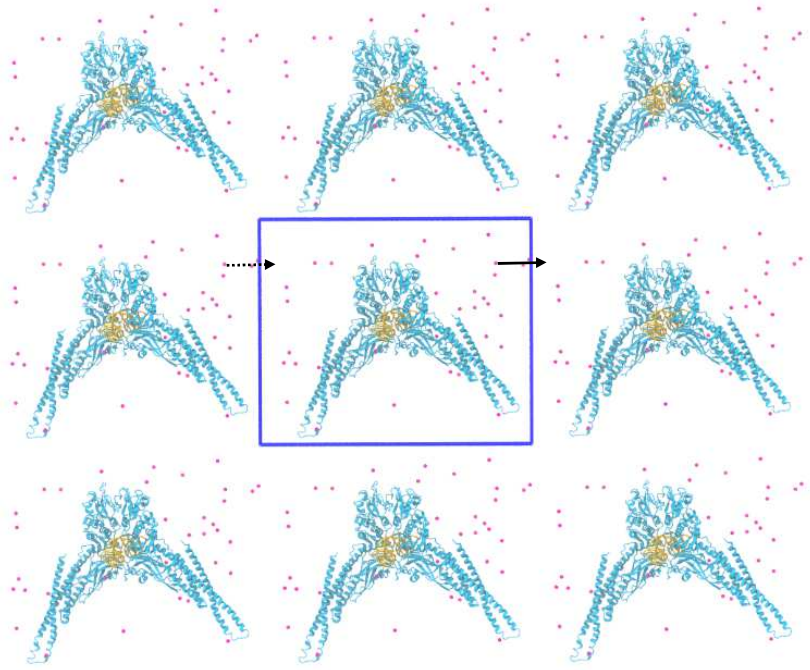


Figure 25: **Conditions périodiques autour d'une boîte contenant un dimère de STAT5b lié à l'ADN.** Si un atome sort de la boîte (flèche pleine), une image va rentrer par le côté opposé (flèche en pointillé). La boîte centrale est colorée en bleue, STAT5b est en cyan, l'ADN est en jaune et les ions en magenta. Les molécules d'eau et les images périodiques dans la profondeur ne sont pas représentées pour plus de clarté.

D. Équilibration et production des simulations de dynamique moléculaire

L'ajout des molécules d'eau et d'ions crée de nouveaux artefacts énergétiques ou contacts potentiels. Une nouvelle minimisation du système est donc requise afin d'obtenir un système sans contraintes, adapté à son environnement aqueux. Trois étapes successives ont été adoptées afin d'atteindre un minimum énergétique du système :

- dans un premier temps, des contraintes de position sont appliquées sur tous les atomes lourds (tous les atomes non-hydrogène) des fragments protéiques ou nucléiques. 10 000 pas de minimisation sont faits, en utilisant l'algorithme de gradient conjugué.
- 10 000 nouveaux pas de gradient conjugué sont à nouveau réalisés, en appliquant des contraintes de positions sur les carbones α des protéines ou sur les atomes de phosphore de la chaîne principale de l'ADN.
- Enfin, 10 000 pas de gradient conjugué sont réalisés sans contrainte.

À partir de ces systèmes minimisés, une impulsion est donnée afin de démarrer une simulation de dynamique moléculaire, les vitesses de départ étant déterminées à l'aide d'une distribution de Boltzmann (*cf.* équation 6, paragraphe II.D.2.a) du chapitre 1). Toutes les simulations de dynamique moléculaire ont été réalisées en utilisant l'intégrateur « en saut de grenouille », en appliquant les conditions périodiques dans les trois dimensions. Pour les

interactions de Van der Waals et électrostatiques, nous avons utilisé une distance seuil de troncation de 1,2 Å, et l'algorithme *Particle Mesh Ewald* pour le calcul des forces électrostatiques longue-distance (*cf.* paragraphe II.D.2.a) du chapitre 1).

Le contrôle de la température est assuré par un couplage à un thermostat de Berendsen utilisant un terme stochastique appliqué aux vitesses afin de générer un ensemble canonique correct[548,549]. Afin de limiter l'apport brusque d'énergie et de ne pas risquer d'endommager la structure, les systèmes simulés ont été chauffés progressivement de 0 à 310K au cours d'une simulation de 100 picosecondes (ps) en appliquant des contraintes sur les atomes de carbone α des protéines et les atomes de phosphore des chaînes principales de l'ADN. Une première simulation de dynamique moléculaire sans contraintes pendant 100 picosecondes (ps) a ensuite été générée dans l'ensemble NVT, avant 100 ps dans l'ensemble NPT en réalisant un couplage de la pression à 1 bar en utilisant l'algorithme de Parinello et Rahman[550]. Pour ces deux dernières simulations et les suivantes, les liaisons impliquant des atomes d'hydrogène ont été traitées par l'algorithme LINCS, ce qui nous a permis d'utiliser un pas d'intégration de 2 femtosecondes (fs). Enfin, une simulation d'équilibration de 5 nanosecondes (ns) a été réalisée pour chaque système dans ces mêmes conditions, à 310K et 1 bar. Les modèles obtenus à la fin de ces étapes d'équilibration constituent les modèles initiaux pour les simulations de dynamique moléculaire de production.

Les mêmes conditions ont été conservées pour les simulations de production des données, mais un nouveau jeu de vitesses initiales a été généré pour chaque simulation. Chaque protéine monomérique, STAT5a non-phosphorylé (STAT5a), STAT5a phosphorylé (pSTAT5a), STAT5b non-phosphorylé (STAT5b) et STAT5b phosphorylé (pSTAT5b) a été simulée pendant 30 ns, au cours de deux réplicas indépendants. Une simulation de STAT5a et une simulation de STAT5b ont été prolongées jusqu'à 200 ns. Chaque protéine dimérique liée à l'ADN, STAT5a porteur d'une histidine 471 protonée en position δ (dSTAT5a^{HID} et dSTAT5b^{HID}), en position ε (dSTAT5a^{HIE} et dSTAT5b^{HIE}) ou doublement protonée (dSTAT5a$_{HIP}$ et dSTAT5b^{HIP}), ont été simulés pendant 30 ns sur un réplica. La durée totale des simulations de STAT5 est donc de 760 ns.

E. Analyses des simulations de dynamique moléculaire

Les trajectoires générées à partir des simulations de dynamique moléculaire constituent la base de données de laquelle nous pouvons extraire des valeurs quantitatives et qualitatives caractérisant la structure et/ou son comportement dynamique. Parmi ces analyses, les plus courantes sont les mesures de déviation des structures générées par rapport à une structure de référence, l'analyse des fluctuations atomiques par résidu et la détection de liaisons hydrogènes observées au cours des simulations.

1. *Mesure des déviations*

L'analyse des déviations structurales est une méthode standard de détection de mouvements de grande amplitude. Pour chaque conformation des trajectoires de dynamique moléculaire, la déviation (*Root Mean Square Deviation, RMSD*) est calculée par rapport à la structure de référence, à savoir la structure initiale de la simulation, au temps t = 0 ns, ou la structure moyenne calculée à partir de l'ensemble des conformations.

$$RMSD\left(t, t^{ref}\right) = \sqrt{\frac{1}{M}\sum_{i=1}^{N} m_i |r_i(t) - r_i(t^{ref})|^2}\,. \qquad \textbf{Équation 22}$$

Cette mesure quadratique de la distance permet une première validation de la stabilité du système au cours de la dynamique, de telle sorte que si le système étudié dérive de manière significative, une visualisation de la dynamique doit révéler un mouvement de grande amplitude. Les structures analysées $r_i(t)$ et de références $r_i(t^{ref})$ sont préalablement superposées afin de ne pas prendre en compte les mouvements de rotation et de translation. Par ailleurs, les fluctuations atomiques moyennes peuvent être calculées sur l'ensemble des conformations générées par rapport à leur position moyenne selon :

$$RMSF(i) = \sqrt{\frac{1}{N}\sum_{t=1}^{T} |r_i(t) - \langle r_i \rangle|^2}\,. \qquad \textbf{Équation 23}$$

Ces calculs peuvent être faits pour l'ensemble des atomes constituant le système. Les mouvements des chaînes latérales peuvent être cependant très supérieurs à ceux des atomes de la chaîne principale, du fait de la présence de liaisons rotatives. Afin de ne pas introduire de bruits dans ces analyses, elles ne sont effectuées que sur les atomes de carbone α pour les résidus protéiques et les atomes de phosphore pour les résidus nucléotidiques. Ces calculs ont été réalisés par les fonctions g_rms et g_rmsf du logiciel GROMACS[539] pour les calculs de déviation et de fluctuations atomiques, respectivement.

2. *Détection des liaisons hydrogènes inter-molécules*

Les liaisons hydrogènes sont l'un des éléments les plus importants des interactions non-covalentes entre macromolécules; la détection de ces liaisons est un enjeu majeur dans l'analyse des interactions au sein des complexes macromoléculaires. Les liaisons hydrogènes dans les dimères de STAT5 liés à l'ADN ont été calculées selon des critères géométriques. Soient A un atome accepteur de liaison hydrogène, D un atome donneur d'hydrogène et H l'atome d'hydrogène entre les atomes D et A. Une liaison hydrogène sera considérée comme existante si la distance D-A est inférieure ou égale à 3,6 Å et si l'angle A-D-H est inférieur ou égal à 30°. Ces calculs ont été faits par la fonction *g_hbond* du logiciel GROMACS[539].

La détection des ponts d'eau n'est pas incluse dans la fonction *g_hbond*. Or il a été montré qu'ils peuvent jouer un rôle important dans les liaisons intermoléculaires[551–553]. De plus, une précédente étude de la dynamique du complexe STAT3-ADN dimérique a révélé des zones

à haute densité d'eau à l'interface protéine-ADN[538]. Afin de détecter ces ponts à l'interface protéine-ADN et entre les monomères des STAT5 dimériques, j'ai développé un script Perl (*cf.* Annexe B.) qui permet l'analyse des fichiers de sortie de GROMACS. Ce script permet, à partir des données qui décrivent les interactions aux interfaces *protéine - eau* et *eau - ADN*, de caractériser les paramètres des ponts d'eau liants des molécules biologiques (A et B) : **protéine – eau - ADN** ou **protéine – eau – protéine,** où A et B sont des protéines ou de l'ADN**.** Les fichiers de sortie décrivent les liaisons hydrogène *molécule A - eau* et *eau – molécule B*, et les ponts *molécule A – eau – molécule B.* Les détails sont placés dans deux types de fichiers : un fichier de données qui regroupe les noms des résidus et des atomes donneurs et accepteurs de liaisons hydrogène (pour les couple *molécule A – eau* et *molécule B – eau*) ainsi que le nom de la molécule d'eau quand nécessaire, en plus du temps d'existence de la liaison hydrogène ou du pont d'eau (en pourcentage) au cours de la simulation de dynamique moléculaire. De plus, un fichier image XPM est produit avec la carte d'existence de chaque pont d'eau.

En l'état actuel, ce script présente plusieurs défauts non négligeables. En particulier, les fichiers d'entrée et de sortie étant très volumineux (plusieurs giga-octets, Go), le script nécessite une grande quantité de mémoire, les données étant chargées dans la mémoire vive de l'ordinateur pour pouvoir être traitées. De plus, parcourir des fichiers de plusieurs millions de lignes à de nombreuses reprises reste un processus coûteux en termes de temps de calcul. L'implémentation du parallélisme a été réalisée pour les fonctions les plus longues, le nombre de processeurs à utiliser étant laissé à l'appréciation de l'utilisateur, mais conduit à une augmentation des besoins en mémoire vive. Un enjeu est donc de trouver un bon équilibre entre mémoire disponible et temps de calcul. Dans le cas des dimères de STAT5, plusieurs jours de calculs sont nécessaires pour chaque molécule lorsqu'une parallélisation sur deux processeurs est faite. Plusieurs pistes de travail sont à l'étude afin de limiter ce phénomène : (i) ne plus charger les données dans la mémoire vive, (ii) limiter les copies de données au cours de la phase de parallélisation, et (iii) parcourir moins souvent les fichiers afin de limiter le temps de calcul. Par ailleurs, dans l'optique d'optimiser les ressources nécessaires, les calculs de liaison hydrogène de type *protéine – eau* ont été limités aux domaines particuliers présentant un intérêt pour cette étude, comme le domaine de liaison à l'ADN dans le cas des ponts ***protéine – eau – ADN***, et réalisés indépendamment pour chaque monomère de STAT5.

3. Rayon de courbure des hélices α

Le rayon de courbure au cours du temps de simulation de dynamique moléculaire a été calculé grâce à l'extension Bendix[554] implémentée dans VMD[446]. La courbure locale de l'hélice est calculée pour chaque résidu d'une hélice et notée dans un fichier texte.

4. Cartes volumétriques des molécules d'eau

Pour chaque simulation de dynamique moléculaire des STAT5 dimériques, nous avons générés des cartes de distribution d'eau. Pour chaque conformation issue de la simulation de

DM, chaque point de la grille se voit attribuer 1 lorsqu'il est occupé par une molécule d'eau et 0 sinon. La moyenne arithmétique de chaque point est calculée et placée dans un fichier que l'on peut visualiser par la suite. La superposition des conformations avant ces calculs est nécessaire pour éviter les artefacts liés aux mouvements de rotation et translation du système simulé. Ces cartes ont été générées par l'extension *VolMap* de VMD[446].

F. Caractérisation de la Dynamique Essentielle de STAT5

1. *Analyse en Composante Principale (ACP)*

Les trajectoires de dynamiques moléculaires ont également été caractérisées par dynamique essentielle, également appelée analyse en composante principale, à l'aide du logiciel ProDy[555]. La covariance des fluctuations des positions atomiques observées au cours de la dynamique est analysée, permettant ainsi l'analyse des mouvements anharmoniques, à l'inverse des modes normaux (*cf.* paragraphe II.F du chapitre 2).

La matrice de covariance C_{ij}est construite à partir des mouvements atomiques autour de leurs positions moyennes déterminées par la trajectoire:

$$C_{ij} = \langle (r_i - \langle r_i \rangle)\big(r_j - \langle r_j \rangle\big) \rangle \qquad \textit{Équation 24}$$

où C_{ij} est l'élément de la matrice correspondant aux atomes i et j, r_i et r_j sont les coordonnées cartésiennes des atomes i et j, respectivement.

Au cours de ce travail, seuls les carbones α (Cα) et les atomes de phosphore (P) de la chaîne principale de l'ADN ont été analysés. Les crochets ⟨ ⟩ indiquent une moyenne mesurée sur l'ensemble des conformations générées au cours de la simulation DM. Chaque conformation de STAT5 est superposée sur la *structure moyenne* afin d'éliminer les mouvements de translations et de rotations. La *structure moyenne* est calculée pour chaque trajectoire à partir de l'ensemble des conformations, et chaque conformation de la trajectoire est superposée à la *structure moyenne* par minimisation du RMSD calculé sur les atomes de Cα uniquement. La matrice symétrique C peut être diagonalisée par l'opération:

$$A^T C A = \lambda \qquad \textbf{Équation 25}$$

où A représente les vecteurs propres de la matrice de covariance et λ_i la valeur propre associée au $i^{ème}$ vecteur propre de la matrice A.

Les valeurs propres définissent la fluctuation totale du système le long du vecteur propre associé[556]. Le nombre de vecteurs propres générés est de l'ordre de 3N – 6, où N correspond au nombre d'atomes considérés.

2. Corrélations croisées

Les corrélations entre les déplacements atomiques du système simulé sont mesurées en calculant les corrélations croisées CC_{ij}^{ACP}. La matrice des corrélations croisées entre chaque paire d'atome est construite par l'équation suivante :

$$CC_{ij}^{ACP} = \frac{\langle \Delta r_i \Delta r_j \rangle}{\langle \Delta r_i^2 \rangle^{1/2} \langle \Delta r_j^2 \rangle^{1/2}} \qquad \textbf{Équation 26}$$

où i et j sont deux carbones α; Δr_i et Δr_j sont les déplacements des atomes i et j par rapport à leur position moyenne et les crochets $\langle\ \rangle$ représentent une moyenne sur l'ensemble des conformations d'une trajectoire de dynamique moléculaire.

Lorsque les mouvements des atomes i et j sont corrélés, l'élément CC_{ij}^{ACP} sera proche de 1 ; lorsque ces mouvements sont décorrélés, $CC_{ij}^{ACP} = 0$; et une anti-corrélation sera donnée par une valeur proche de -1.

L'ensemble des calculs de l'analyse en composante principale a été réalisée à l'aide du logiciel ProDy[555].

III. Modes Normaux calculés avec un modèle en réseau anisotrope

A. Calcul de la matrice hessienne et équations dérivées

L'analyse des modes normaux (ANM) a été réalisée à partir de plusieurs structures : les modèles par homologie générés par Modeller, les modèles par homologie minimisés énergétiquement, les modèles du système chauffés et équilibrés, et les conformations finales des trajectoires de dynamique moléculaire. Les résultats d'analyse obtenus sont très similaires pour chaque système. Par conséquent, nous ne présenterons que les résultats obtenus à partir des conformations équilibrées, qui correspondent au temps initial (t = 0 ns) des simulations de production. Toutes les analyses ont été réalisées avec ProDy[555].

Nous avons utilisé le modèle en réseau anisotrope, un réseau élastique développé initialement dans le groupe d'I. Bahar[494,557]. Dans ce modèle, les résidus sont connectés entre eux par des élastiques dont la constante de force k varie en fonction de la connectivité des résidus (*cf.* paragraphe II.E du chapitre 1). Dans les modèles à réseau anisotrope, un facteur de connectivité est ajouté (*cf.* équation 20 du chapitre 1). Les facteurs de connectivité sont stockés dans une matrice de Kirchhoff Γ, les éléments $\Gamma_{a,b}$ contenant la connectivité entre les atomes a et b. En intégrant la matrice de Kirchhoff avec les équations 20 et 21 du chapitre 1, on obtient la formulation suivante :

$$V = \frac{\mathrm{k}}{2}\sum_{a,b}^{M}\left(\Gamma_{a,b}\right)\left(\left|r_{a,b}\right| - \left|r_{a,b}^{0}\right|\right)^{2} \qquad \textbf{Équation 27}$$

où M est le nombre de ressorts.

La matrice hessienne, de dimension $3\mathrm{N} \times 3\mathrm{N}$ associée à cette fonction d'énergie peut être écrite sous la forme d'une matrice composée de $\mathrm{N} \times \mathrm{N}$ éléments de la forme :

$$\mathrm{H}_{\mathrm{ab,b\neq a}} = \frac{\gamma\Gamma_{\mathrm{a,b}}}{\left(\mathrm{R}_{\mathrm{a,b}}^{0}\right)^{2}}\begin{bmatrix}\mathrm{X_{ab}X_{ab}} & \mathrm{X_{ab}Y_{ab}} & \mathrm{X_{ab}Z_{ab}}\\ \mathrm{Y_{ab}X_{ab}} & \mathrm{Y_{ab}Y_{ab}} & \mathrm{Y_{ab}Z_{ab}}\\ \mathrm{Z_{ab}X_{ab}} & \mathrm{Z_{ab}Y_{ab}} & \mathrm{Z_{ab}Z_{ab}}\end{bmatrix}, \qquad \textbf{Équation 28}$$

$$\text{et } \mathrm{H_{ab}} = -\sum_{\mathrm{b,b\neq a}}\mathrm{H_{ab}}. \qquad \textbf{Équation } \textit{29}$$

Cette matrice hessienne permet d'obtenir 3N vecteurs propres u_k et leurs valeurs propres associées λ_k. Les six derniers modes sont cependant associés à des valeurs propres nulles et correspondent aux mouvements de translation et rotation. Les premiers modes décrivent des mouvements amples et concertés de résidus, et correspondent à des déplacements de basses fréquences généralement associés aux mouvements fonctionnels des molécules biologiques [558].

Pour élaborer les modèles anisotropes des différentes espèces de STAT5, nous avons connecté les carbones α des résidus protéiques ou les atomes P, C1' et C2 des acides nucléiques (*cf.* Figure 26). Ces atomes constituent les nœuds du réseau, et sont reliés par des ressorts.

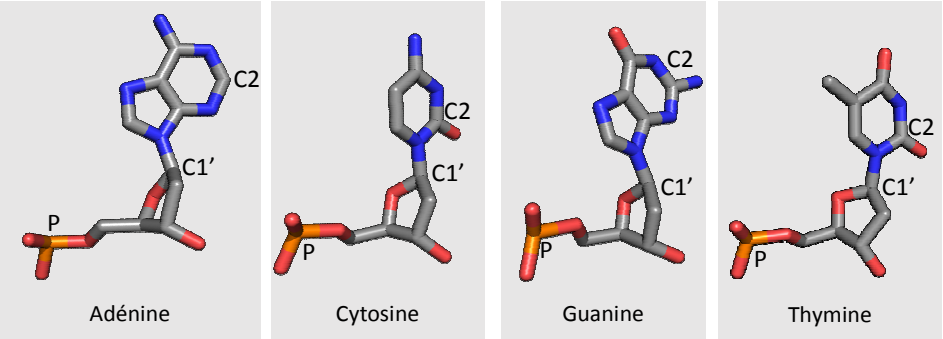


Figure 26: **Emplacement des atomes P, C1' et C2 dans les résidus nucléiques.**

Les constantes de forces des ressorts entre les paires d'atome ont été attribuées en fonction de la proximité spatiale des nœuds, mais également en fonction de la présence des résidus dans les mêmes structures secondaires. Cette approche permet une meilleure description du réseau élastique des structures protéiques, en termes de déplacements des résidus par rapport à leurs voisins proches[559]. Les brins d'ADN complémentaires ont été traités comme un feuillet β. Les structures secondaires ont été déterminées par le logiciel DSSP[560,561] et intégrés dans le fichier de coordonnées à l'aide de l'utilitaire dssp2pdb de James Stroud (http://dssp2pdb.bravais.net/).

Ainsi, les forces suivent le schéma suivant :

- $k = 10$ pour les résidus successifs (résidus n et $n + 1$),
- $k = 6$ pour les nœuds distants de moins de 7 Å et appartenant à la même hélice, et pour les nœuds distants de moins de 6 Å et appartenant au même feuillet β,
- $k = 1$ pour les nœuds distants de moins de 10 Å,
- $k = 0$ dans les autres cas.

De même que pour les analyses en composante principale, les corrélations croisées entre les déplacements des résidus peuvent être calculées à partir de la matrice de covariance. Dans le cadre des modèles à réseau anisotrope, la matrice de covariance est construite à partir d'une matrice pseudo-inverse, notée H^{-1} de la matrice hessienne H :

$$H^{-1} = \sum_{i=1}^{3N-6} \frac{1}{\lambda_i} u_i u_i^T \qquad \textbf{Équation 30}$$

où u_i est le $i^{ème}$ vecteur propre de la matrice hessienne.

La matrice pseudo-inverse est également organisée en sous-matrices de taille 3×3. Les corrélations croisées entre les atomes i et j sont ensuite données par la formule :

$$CC_{ij}^{AMN} = \frac{tr\left(H_{ij}^{-1}\right)}{\sqrt{tr\left(H_{ii}^{-1}\right) \cdot tr\left(H_{jj}^{-1}\right)}} \qquad \text{Équation 31}$$

où tr est la trace de la sous-matrice.

B. Avantages et limitations d'un modèle élastique anisotrope

D'autres types de modèles, décrits brièvement dans le paragraphe (*cf.* paragraphe II.E du chapitre 1), sont disponibles pour le calcul des modes normaux. Le choix d'un réseau élastique a été principalement motivé par la faible sensibilité de cette méthode à la position de la molécule sur sa surface énergétique à proximité d'un minimum énergétique global. Cette propriété est particulièrement intéressante dans le cas de larges macromolécules et de complexes, pour lesquelles parvenir à une structure minimisée au sens énergétique peut poser un problème à la fois en temps de calcul mais également en termes de représentativité de la conformation analysée. Un réseau élastique offre de ce point de vue davantage de flexibilité lorsque peu de données structurales sont disponibles, comme dans le cas des protéines STAT5s.

Les modèles en réseau anisotrope présentent également l'avantage de se placer dans un espace à trois dimensions. Ils apportent donc des informations relatives à la direction des mouvements, à l'inverse d'autres modèles tel que le modèle en réseau gaussien (*Gaussian Network Model, GNM*), qui se placent dans un espace à une seule dimension[493]. Les fluctuations des résidus sont isotropes dans ce cas : les composantes x, y et z ne peuvent pas être isolées à partir des modes normaux. Les amplitudes des déplacements atomiques ou résiduelles sont cependant plus fiables lorsqu'on compare les résultats des deux méthodes, ANM et GNM, aux résultats expérimentaux présentant les facteurs thermiques isotropes (B_{iso}) [494,557].

Pour conclure, les modèles en réseau isotropes comme le modèle en réseau gaussien GNM, sont plus précis et permettent d'évaluer avec une plus grande précision les points de déformation d'une structure. Les modèles anisotropes permettent eux la prédiction des directions des mouvements, bien que l'évaluation des déformations soit légèrement moins réaliste. L'utilisation de tels réseaux est néanmoins nécessaire pour étudier les mécanismes des mouvements d'une protéine.

IV. MOdular NETwork Analysis - MONETA

A. Méthodes bioinformatiques d'analyse des réseaux allostériques

L'allostérie est un phénomène biologique difficile à mesurer expérimentalement et/ou à décrire par des méthodes théoriques à l'échelle atomique. Le développement de méthodes expérimentales permettant d'observer certaines manifestations de ce phénomène à l'échelle quasi-atomique et l'application d'approches théoriques innovantes ont offerts de nouvelles bases pour la modélisation *in silico* des réseaux allostériques au sein des macromolécules biologiques et de leurs complexes. Le point commun de ces méthodes est l'utilisation des structures disponibles, et/ou des simulations de dynamique moléculaire associées à ces structures, au niveau atomique, et de descripteurs géométrique ou énergétique, ou une combinaison des deux. Les protéines allostériques sont donc des **commutateurs d'information**: les signaux apportés par des *stimuli* (effecteurs) sont détectés, intégrés et transmis *via* des voies de communication à travers la protéine en des sites où est élaborée une réponse spécifique. Il y aurait plusieurs types de mécanismes allostériques, sous-tendus par différents liens de causalité entre repliement/rigidité et flexibilité des sites effecteurs, des sites affectés et des segments les liant. Cette vision de l'allostérie comme un phénomène global à l'échelle de la protéine ne contredit pas l'idée de transmission de proche en proche de perturbations locales.

Ainsi, plusieurs tentatives de représentation de la transmission d'information allostérique entre résidus peuvent être citées. La représentation des réseaux allostériques peut être basée sur la description du paysage structural autour de chaque réseau de la protéine: deux résidus vont pouvoir transmettre une information seulement si ils se contactent (liaisons non-covalentes). Ces observations peuvent être extraites soit de structures cristallographiques [562–564] soit de simulations de dynamique moléculaire[565–567].

Chennubhotla et Bahar[563] ont présenté une approche basée sur l'analyse du nombre d'interactions entre paire d'atomes d'une protéine et qui débouche sur la notion de temps de trajet entre paires d'atomes. Cette quantité représente la probabilité pour un signal de transiter entre la paire de résidu considérée. De manière similaire, Atilgan et collaborateurs [562] ont associé à chaque paire de résidus proches un potentiel d'interaction inter-résidus de type énergétique. Plusieurs descripteurs (connectivité moyenne, longueur du plus court chemin,…) ont été présentés afin de caractériser certaines propriétés du réseau d'interaction. Park et Kim[564] ont par la suite utilisé la définition des contacts inter-résidus développé par Chennubhotla et Bahar[563] et l'ont enrichi par une chaîne de Markov afin (i) de générer des temps de visite attendus qui représente une quantification du trajet d'un signal en analysant l'ensemble des trajets possibles, et (ii) de détecter des sites aux interfaces des résidus qui pourraient jouer un rôle important pour la transmission d'un signal. Ces méthodes ont été développées à partir de structures figées, issues de la cristallographie par rayons X notamment, et ne prennent pas en compte les effets associés à la dynamique des protéines.

Afin de prendre en compte les effets dynamiques, Sethi et collaborateurs ont utilisé les probabilités de contact de trajectoires de dynamique moléculaire afin de pondérer les sommets du graphe d'interaction du système étudié[566]. Les poids associés aux interactions inter-résidus individuels servent ensuite à proposer des chemins entre sites distants d'une protéine, qui peuvent être analysés en fonction de leurs longueurs, alors que les nœuds (les résidus du système) se voient affecter un score en fonction des chemins les traversant. Une représentation des « communautés » de résidus hautement connectés entre eux est effectuée et les résidus clé entre ces communautés sont proposés comme ayant un rôle important dans l'allostérie. Enfin, une approche alternative a été présentée récemment, qui porte sur l'analyse entropique de systèmes allostériques complexes (protéines intrinsèquement désordonnées) afin de prédire l'affinité de liaisons protéine-protéine[567] à partir de descripteurs internes. Enfin, ces méthodes restent peu adaptées à l'évaluation des transitions allostériques entre différents états macromoléculaires bien que des méthodes aient été proposées[565].

L'analyse des réseaux modulaires (MOdular NETwork Analysis, MONETA) a été développée dans le contexte de l'analyse des réseaux allostériques intra-protéiques dans des objets relevant du domaine de l'oncologie, et plus particulièrement dans le but de décrire et prédire les effets de mutations oncogéniques, nombreuses dans ce champ d'étude. L'étude initiale était l'analyse des phénomènes d'activation/désactivation des récepteurs tyrosine kinases (RTKs) affectés par des mutations ponctuelles distantes des sites fonctionnels (site de liaison à l'ATP, sites de phosphorylation)[568–570]. L'apparition de mutations équivalentes au sein de ces protéines (par exemple la mutation D802V du CSF-1R et D816V de KIT) engendre des effets structuraux et dynamiques différents. Leur analyse *via* MONETA a permis d'émettre de nouvelles hypothèses quant à l'activation constitutive des formes mutées ou/et à la résistance aux traitements médicamenteux actuels.

MONETA[568,571] se base sur les paramètres issus de l'analyse statistique de simulations de dynamique moléculaire tout-atome de manière à prendre en considération les variations tant de géométrie que de contacts inter-résidus, effets du solvant, *etc.* (*cf.* Figure 27).

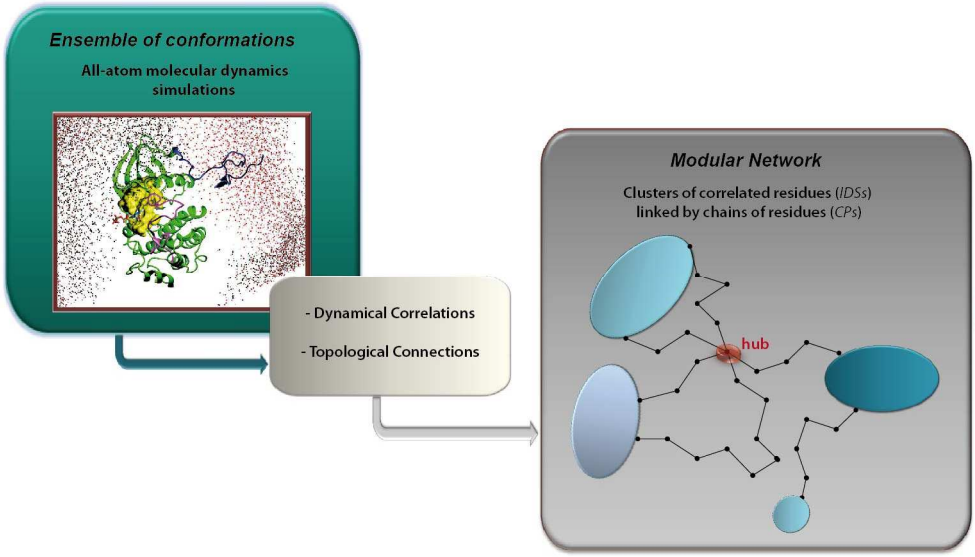


Figure 27: **Représentation schématique des entrées et sorties de MONETA.**

Partant d'une ou plusieurs simulations, trois étapes successives sont réalisées :

- La recherche de Segments Dynamiques Indépendants (*Independent Dynamics Segments*, *IDS*s) représentant des groupes de résidus dont la dynamique interne est très concertée mais très décorrélée du reste du système;
- Les chemins de communication (*Communication Pathways*, *CP*s) sont déduits des contacts inter-résidus observés au cours des simulations de dynamique moléculaire ;
- La visualisation des *CP*s peut se faire en deux dimensions (analyse globale et locale du réseau de communication) ou en trois dimensions (*IDS*s et *CP*s individuels), de manière interactive (Figure 28).

Les deux premières étapes et leurs concepts théoriques sous-jacents sont détaillés de manière explicite dans les prochains paragraphes, ainsi que les approches en développement.

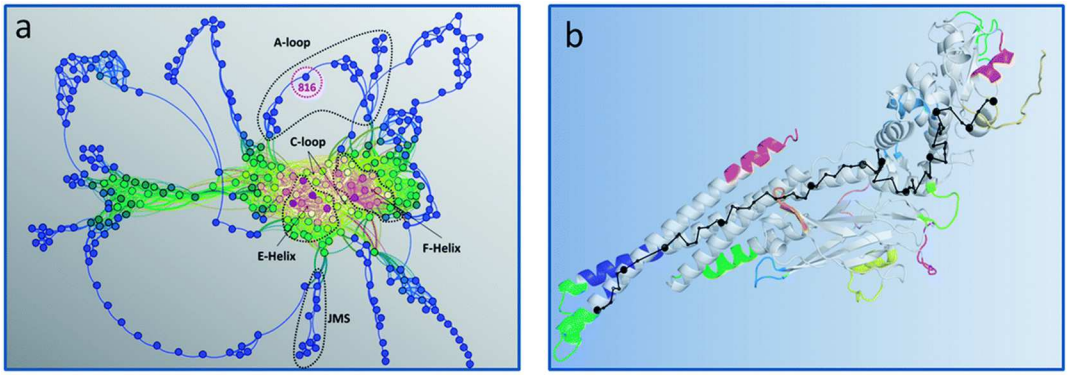


Figure 28 : **Représentations des sorties de MONETA.** (a) Réseau des chemins de communication dans KIT, un récepteur tyrosine kinase de type III. Les résidus de certains éléments fonctionnels (A-loop, E-helix, …) sont entourés en pointillés. (b) Chaque IDS de STAT5a est représenté en couleur, et une succession de chemins de communication reliant deux sites distants est représentée en noir. Figure reproduite avec la permission des auteurs[571].

B. *Independent Dynamics Segments*, IDSs, et notions théoriques associées

Au cours du développement initial de MONETA, l'indentification des Segments Dynamiques Indépendants (*Independent Dynamic Segments*, *IDS*s) a été réalisée à partir de l'approche statistique dite de l' « analyse des traits locaux » (*Local Feature Analysis, LFA*)[572]. Initialement développé pour traiter les données d'analyses en composante principale (ACP) dans le cadre d'analyse d'images, elle a été par la suite adaptée[573,574] pour traiter les résultats issus de trajectoires de dynamique moléculaire. L'ACP permet une réduction de la dimensionnalité de données, par exemple issues des simulations de dynamique moléculaire. Le formalisme *LFA* appliqué ici est une méthode particulière permettant de traiter les résultats d'ACP et destinée à fournir une représentation topographique de la dynamique d'une protéine qui est localisée, à l'inverse de l'ACP où la dynamique est largement distribuée. Cependant, à l'inverse de Zhang & Wriggers, nous n'étendons pas la description topographique à l'ensemble de la protéine[568], comme décrit ci-dessous.

Soient m le nombre de conformations générées au cours d'une simulation et N le nombre d'atomes dans le système simulé par dynamique moléculaire. L'analyse en composante principale des déplacements atomiques produit 3N modes principaux orthogonaux à partir desquelles peuvent être reconstruites les composantes des coordonnées à partir de l'équation 32 :

$$x_i = \sum_{k=1}^{3N} A_k \psi_k(i) \,, \qquad \textbf{Équation 32}$$

$$\text{avec } A_r = \sum_{i=1}^{3N} \psi_k(i)\, x_i \equiv \sum_{i=1}^{3N} K_r^{ACP}\,(i) x_i \,, \qquad \textbf{Équation 33}$$

où A_r est la projection des déplacements atomiques sur le mode principal ψ_k, et K_r^{ACP} est appelé noyau de l'ACP.

Les modes principaux étant classés par décroissant de leurs valeurs propres, les n premiers modes principaux, avec $n \ll 3N$, peuvent être utilisés pour reconstruire une trajectoire représentant la dynamique essentielle du système :

$$x_i' = \sum_{k=1}^{n} A_k \psi_k(i) \,. \qquad \textbf{Équation 34}$$

Les calculs d'ACP ont été faits pour tous les systèmes étudiés à partir des coordonnées atomiques des atomes de carbone α, en concaténant les trajectoires de chaque paire de réplica pour les STAT5 monomériques. Les vecteurs issus de ces calculs ont été utilisés pour appliquer le formalisme d' « analyse des traits locaux ». Pour STAT5a, pSTAT5a, STAT5b et pSTAT5b, nous avons conservé $n = 41$, 41, 15 et 48 modes principaux, ce qui correspond au nombre de modes nécessaires pour décrire 95% des fluctuations atomiques des systèmes. Ici, n correspond ainsi aux 2,41, 2,41, 0,87 et 2, 79% premiers modes principaux, ce qui offre une importante réduction de la dimensionnalité des données. Mais ces données sont globales, le noyau s'applique donc à l'ensemble des résidus i, les k modes étant indépendants car orthogonaux. Le formalisme *LFA*

a pour but de redistribuer l'information par résidu i plutôt que par mode r. Pour transformer l'information globale en information locale, le noyau *LFA* prend la forme :

$$K^{LFA}(i,j) = \sum_{k,l}^{n} \psi_k(i) Q_{kl} \psi_l(j) \,, \qquad \textbf{Équation 35}$$

avec Q_{kl} une matrice arbitraire réelle. Par similarité avecl'équation 33, les modes *LFA* définissent des projections locales qui ne dépendent plus du mode r mais du résidu j :

$$\mathrm{O(j)} \equiv \sum_{\mathrm{j=1}}^{\mathrm{3N}} \mathrm{K^{LFA}\,(i,j) x_j} = \sum_{\mathrm{k,l}}^{\mathrm{n}} \psi_{\mathrm{k}}(\mathrm{j}) \mathrm{Q_{kl} A_k} \,. \qquad \textbf{Équation 36}$$

Les $\mathrm{3N}$ projections ne peuvent être entièrement décorrélées puisque construits à partir d'un jeu de seulement $\mathrm{n} \ll \mathrm{3N}$ modes. Afin d'assurer une décorrélation maximale des $\mathrm{O_j}$, Penev et Attick ont montré que le noyau LFA est de la forme[572] :

$$\mathrm{K^{LFA}(i,j)} = \sum_{\mathrm{k=1}}^{\mathrm{n}} \psi_{\mathrm{r}}(\mathrm{i}) \frac{1}{\sqrt{\lambda_{\mathrm{k}}}} \psi_{\mathrm{r(j)}} \,, \qquad \textbf{Équation 37}$$

d'où, en combinant avec l'équation 36, on obtient : $\mathrm{O(j)} = \sum_{\mathrm{k=1}}^{\mathrm{n}} \frac{\mathrm{A_k}}{\sqrt{\lambda_{\mathrm{k}}}} \psi_{\mathrm{k}}(\mathrm{j}),$ **Équation 38**

dont les corrélations résiduelles dont données par :

$$\langle \mathrm{O(i)O(j)} \rangle = \sum_{\mathrm{k=1}}^{\mathrm{n}} \psi_{\mathrm{k}}(\mathrm{i}) \psi_{\mathrm{k}}(\mathrm{j}) \equiv \mathrm{P(i,j)}. \qquad \textbf{Équation 39}$$

L' « analyse des traits locaux » remplace les n premiers modes principaux globaux par $\mathrm{3N}$ traits locaux $\mathrm{O(j)}$ soit le nombre de degré de liberté de départ. Plutôt que de retourner à un ensemble de dimension $\mathrm{3N}$, un algorithme va sélectionner M traits locaux qui vont le mieux approximer les $\mathrm{3N}$ traits locaux $\mathrm{O(j)}$. À l'image de l'ACP (équation 34), les traits locaux peuvent être reconstruits à partir d'un sous-ensemble $M = n$ de carbone α. Les traits locaux $O'(i)$ sont reconstruits à partir des éléments de M (équation 40), et l'erreur $O^{err}(i) = \langle \|O(i) - O'(i)\|^2 \rangle$ calculée.

$$O'(i) = \sum_{m=1}^{M} a_m(i) O(i_m), \qquad \textbf{Équation 40}$$

$$\text{avec } a_m(i) = \sum_{l=1}^{M} P(i,i_l) P'^{-1}_{lm} \,, \qquad \textbf{Équation 41}$$

où P'^{-1} est l'inverse de la sous-matrice P' de P, et $P'_{lm} \equiv P(i_l, i_m)$.

Le carbone α présentant l'erreur maximale est celui dont la dynamique est la plus décorrélée par rapport à la dynamique décrite par les éléments de l'ensemble M ; il est donc ajouté à M. Par itération, on sélectionne ainsi le sous-ensemble de carbones α qui décrit le mieux la dynamique de la simulation, dans une limite de m atomes, avec $m = n$, le nombre de modes principaux conservés pour l'analyse. Si un carbone α apparaît à plusieurs reprises, ou si deux éléments de M sont situés à moins de 6 résidus de distance dans la séquence peptidique, le résidu apparu en premier sera conservé, le nombre final d'*IDS*s pourra donc être logiquement inférieur à n. Vu la fraction des fluctuations du système expliquée par les éléments de M, il est fréquent de se trouver dans ce cas. Néanmoins, les résidus qui n'ont pas été retenus sont très souvent intégrés dans un *IDS*s lorsque ceux-ci sont agrandis.

Le sous-ensemble de $m \leq n$ atomes contient les résidus ***graines*** des *IDS*s, qui vont être agrandis progressivement. Un processus itératif permet de cribler les résidus du système à partir de chaque graine pour détecter les résidus présentant une bonne corrélation dynamique avec l'ensemble des atomes de l'*IDS* considéré. Pour chaque *IDS* S_m, le résidu k est inclus si la corrélation moyenne de k avec les membres de S_m est supérieure à une valeur seuil P_{cut}, et si au moins un des résidus de S_m est éloigné d'une distance inférieure à d_{cut} :

$$\frac{\sum_{j=1}^{|S_m|} P(k,j)}{|S_m|} \geq P_{cut}, \forall\, j \in S_m\,. \qquad \textbf{Équation 42}$$

Une valeur seuil P_{cut} arbitraire a été choisie de manière à garder 1,2% des corrélations croisées de LFA. La valeur de P_{cut} a été fixée à 0,032 pour STAT5a, pSTAT5a, 0,019 pour STAT5b et 0,035 pour pSTAT5b. Les matrices comprenant la moyenne de la plus petite distance entre chaque paire de résidus a été calculée en utilisant la fonction *g_mdmat* de GROMACS 4.5. Deux résidus sont considérés comme voisins quand la moyenne de la plus petite distance entre eux est inférieure à la valeur seuil d_{cut} de 3,7 Å. La représentation en réseau modulaire de ces protéines a été réalisée avec la dernière implémentation de l'algorithme LFA dans MONETA [571,575].

C. Voies de communication (*Communication Pathways, CP*) et transmission d'information allostérique

La notion de chemins de communications (*Communication Pathways*, *CP*s) se base sur le travail de Chennubhotla et collaborateurs sur la «*propension des résidus à communiquer*» [563], et est particulièrement avantageuse dans le cadre de la conception de l'allostérie décrite dans le paragraphe III du chapitre 1. Les chemins de communication entre tous les résidus protéiques et nucléiques ont été calculés en utilisant l'algorithme décrit par Laine et al. [568].

Les *CP*s sont agrandis progressivement de manière à ce que deux résidus reliés par une interaction non-covalente et que chaque paire de résidus à l'intérieur d'un *CP* donné présentent un temps de trajet (*Commute Time*, *CT*) inférieur à CT_{seuil}. Les interactions non-covalentes ont été détectées pour toutes les trajectoires par l'utilitaire LIGPLOT[576]. Si une interaction non-covalente est observée entre deux résidus pendant plus de 50% du temps de simulations, ces résidus sont considérés comme interagissant de manière stable. Le seuil de communication CT_{seuil} a été fixé à 0,1, les résidus communiquant le plus efficacement sont ainsi connectés avec 6 – 9 % des résidus dans le cas des STAT5 monomériques (soit 34 à 50 résidus environ).

L'analyse statistique des données générées a été effectuée à l'aide du logiciel R[577], alors que la représentation des *IDS*s et *CP*s dans les structures ont été réalisée dans PyMOL[442] à l'aide de fonctions adaptées et incorporées dans MONETA[571]. La visualisation du réseau de *CP*s a été faite à travers l'utilitaire *gephi*[578].

D. Développement de MONETA et contribution des membres de BiMoDyM

MONETA a suivi un développement séquentiel, initié par le Dr. Luba Tchertanov, fondatrice du groupe BiMoDyM, et le Dr. Elodie Laine, alors post-doctorante dans ce groupe, qui a développé les premiers scripts nécessaires aux calculs pour l'analyse en réseau modulaire[568]. Ce travail a principalement été prolongé par Ariane Allain et Yann Guarracino, étudiants en master 1 (2013) et en master 2 (2014). Ariane a (i) regroupé les différents scripts de MONETA en un programme qui génère les données nécessaires au moyen d'une commande, (ii) amélioré le traitement des fichiers temporaires, (iii) optimisé certaines parties du code et (iv) introduit une nouvelle méthode de visualisation avec *gephi*. Yann a par la suite (i) adapté le code aux nouvelles versions des différents outils utilisés dans MONETA (AmberTools, GROMACS, LigPlot), (ii) parallélisé certaines étapes de calcul, (iii) adapté le code pour permettre la prise en charge des molécules d'acides nucléiques et (iv) introduits de nouvelles pièces de code pour le calcul des *IDS*s[571]. Par ailleurs, Isaure Chauvot de Beauchêne, doctorante au sein du groupe (2011-2013), a également participé au développement de MONETA par l'écriture de certaines parties du code, l'encadrement d'Ariane Allain et des discussions concernant l'évolution de Moneta. Enfin, Nicolas Panel, étudiant en master 2 (2014), a écrit un script permettant la projection des coordonnées atomiques d'un système sur un plan, afin de faciliter la visualisation du réseau de chemins de communication par *gephi*. Les étapes successives des calculs de MONETA sont résumées dans la Figure 29.

La recherche de nouvelles pistes pour caractériser la dynamique moléculaire et sa relation avec le transfert du signal allostérique nous a amené à rentrer en collaboration avec les membres du Centre de Mathématiques et de Leurs Applications (CMLA), et plus particulièrement avec le Pr. Alain Trouvé. Ce travail a été mené dans le cadre des projets de l'Institut FARMAN de l'ENS Cachan et associé aux différents stages des étudiants (niveau License 3) du département de Mathématique de l'ENS Cachan. Cette collaboration a permis le développement d'une nouvelle approche pour le calcul des *IDS*s, décrite ci-dessous.

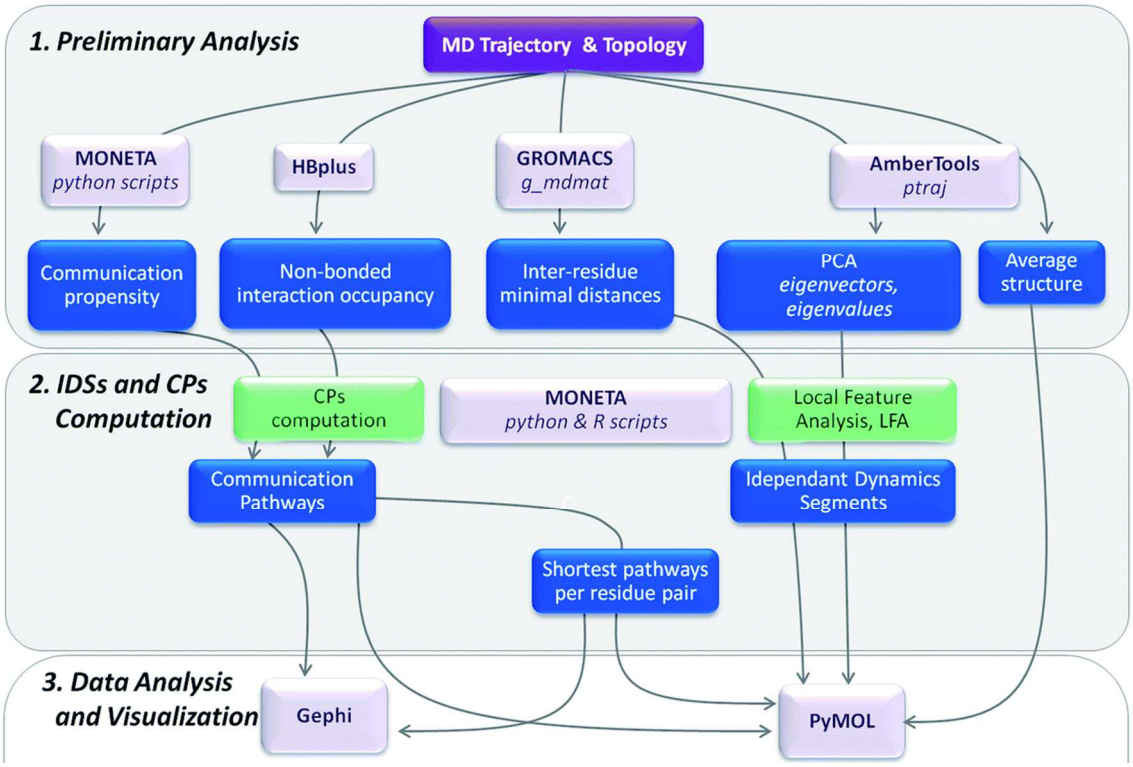


Figure 29: **Les différentes étapes de calcul par MONETA.** (1) Les analyses préliminaires sont réalisées par des utilitaires externes. (2) MONETA regroupe les résultats préliminaires afin de calculer les chemins de communication (*CP*s, à gauche) et les Segments Dynamiques Indépendants (*IDS*s, à droite). (3) La visualisation des résultats peut se faire par *gephi* (analyse du réseau de communication) ou PyMOL (représentation des *IDS*s, et analyse des *CP*s)

E. Développement d'une nouvelle méthode de calcul des *IDSs* : « Décomposition des Traits Principaux »

L'idée principale de l'approche en « décomposition des traits principaux » (*Principal Feature Decomposition, PFD*) est de rechercher de manière récursive les atomes X^j qui peuvent prédire le comportement dynamique d'un sous-ensemble d'atomes (*cf.* Figure 30).

Cependant, les modes lents à l'échelle de l'ensemble du système sont une source importante de corrélation entre des atomes distants, et il est nécessaire de les prendre en compte et de les retirer. Par la suite, nous désignerons $X \in E = \mathbb{R}^{3N}$ le vecteur colonne contenant les coordonnées de tous les carbones α de la protéine, $\overline{X} = \langle X \rangle$ sa moyenne empirique déterminée sur l'ensemble de conformations des trajectoires de dynamique moléculaire et $\Delta X \doteq X - \overline{X}$ le vecteur de déplacement par rapport à la conformation moyenne. ΔX_i indiquera la position centrée du $i^{ème}$ atome du vecteur, et la matrice de covariance est définie par $\Gamma_E \doteq \langle \Delta X \Delta X^T \rangle$.

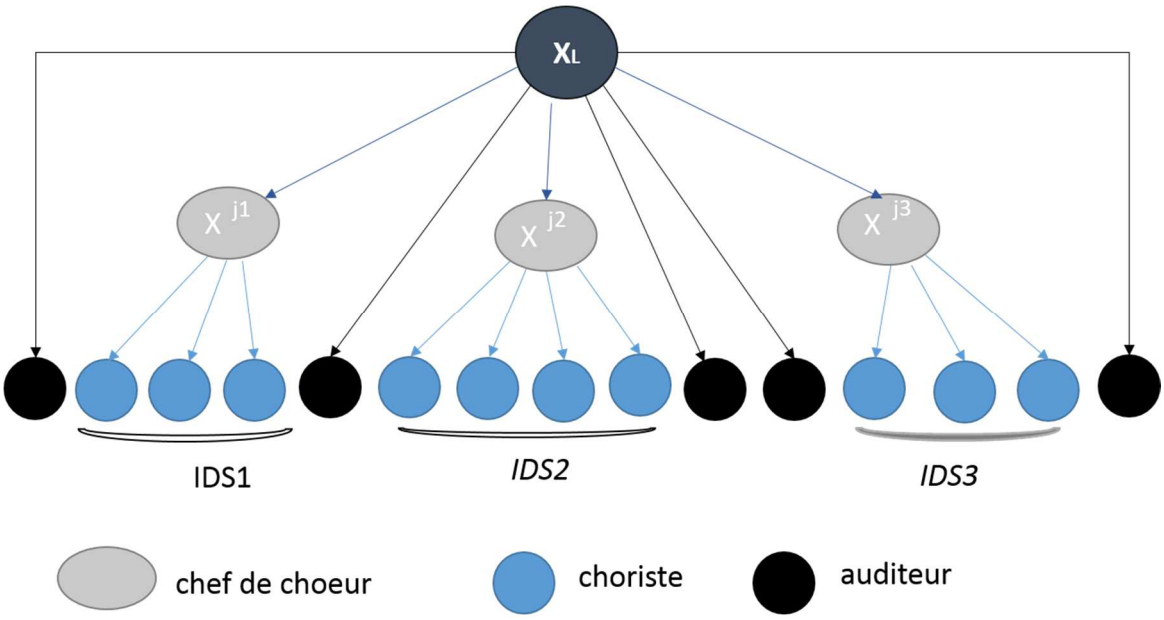


Figure 30: Représentation schématique des IDSs selon l'approche PFD. Plusieurs chefs de chœurs (ou atomes prédicteurs) expliquent une grande part de la dynamique du système global. Des choristes sont ensuite recrutés par les chefs de chœur pour constituer les IDSs.

Modes lents et analyse des corrélations canoniques: la présence de mouvements globaux associés à des fréquences lentes dans les premières valeurs propres de l'ACP de la matrice $\mathbf{\Gamma_E}$ associés aux vecteurs propres $\boldsymbol{\psi_k}$ est une source de corrélation croisée dans la détection des *IDS*s potentiels. L'analyse des corrélations canoniques (ACC) entre les coordonnées de deux atomes i and j est réalisée en calculant :

$$\rho_{ij} \doteq \max_{u,v} \frac{\langle (u,\Delta X_i)(v,\Delta X_j)\rangle}{\sqrt{\langle (u,\Delta X_i)^2\rangle \langle (v,\Delta X_j)^2\rangle}}. \qquad \textbf{Équation 43}$$

Les corrélations canonoiques sont donc telles que $0 \leq \rho_{ij} \leq 1$, et les corrélations canoniques sont toujours supérieures à la valeur absolue des corrélations croisées entre les atomes i and j comme le montre la formule:

$$\rho_{ij} \geq \left| CC_{ij}^{ACP} \right| = \frac{\left|\langle (\Delta X_i,\Delta X_j)\rangle\right|}{\sqrt{\langle |\Delta X_i|^2\rangle \langle |\Delta X_j|^2\rangle}}, \qquad \textbf{Équation 44}$$

dont le terme de droite a été introduit dans l'équation 26. De plus, ρ_{ij} n'est autre que le cosinus du premier angle principal α_{ij} entre les deux espaces $3D$ générés par les trajectoires des trois coordonnées de ΔX_i et ΔX_j.

Nous pensons que le paramètre de corrélation canonique ρ_{ij} est plus pertinent et efficace que le paramètre de corrélation croisée CC_{ij} pour évaluer l'existence d'une dépendance linéaire entre deux atomes, puisque CC_{ij} peut être volatil même si deux vecteurs ΔX_i et ΔX_j sont statistiquement fortement corrélés.

Tableau 4: Valeurs des 1ers et 9èmes deciles des corrélations canoniques ρ_{ij} *(80% des paires* i ≠ j *ont des corrélations comprises entre ces deux valeurs).*

STAT5a	pSTAT5a	STAT5b	pSTAT5b
0,49	0,50	0,77	0,53
0,87	0,85	0,97	0,97

Sur STAT5a, nous avons trouvé que 90% des paires d'atomes, ij , avec $\neq j$, présentent une corrélation canonique $\rho_{ij} \geq 0.49$, et pour STAT5b, cette corrélation est encore plus élevée (0,77, *cf.* Tableau 4). Pour enlever les effets masquant des modes lents (associées aux premières valeurs propres) sur la dynamique locale, nous retirons du vecteur colonne ΔX sa projection $\Delta X^{slow} \doteq \sum_{k=1}^{q}(\psi_k, \Delta X)\psi_k$ sur le sous-espace généré par les q premiers vecteurs propres $(\psi_k)_{1\leq k\leq q}$, de manière à ce que la méthode *PFD* soit appliquée sur la matrice $\Delta X^{(0)} \doteq \Delta X - \Delta X^{slow}$. Les effets de filtration induisent une forte diminution des corrélations ρ_{ij} lorsqu'elles sont calculées sur la matrice $\Delta X^{(0)}$(*cf.* Tableau 5). En effet, si l'on prend la valeur $q = 6$, au moins 90% des paires d'atomes $i \neq j$ ont une corrélation canonique inférieure à 0.64 (ce qui correspond à un angle principal de plus de $50°$).

Tableau 5: Valeurs des 1ers et 9èmes déciles des corrélations canoniques ρ_{ij} après retrait des projections sur les 6 premiers vecteurs propres *(80% des paires* i ≠ j *ont des corrélations comprises entre ces deux valeurs).*

STAT5a	pSTAT5a	STAT5b	pSTAT5b
0,27	0,26	0,23	0,29
0,62	0,61	0,57	0,64

*« **Décomposition des traits principaux » sur les trajectoires filtrées:*** Nous avons trouvé que retirer $q = 6$ vecteurs propres avant d'effectuer les calculs de l'algorithme *PFD* sur la matrice $\boldsymbol{X^{(0)}}$ permet d'obtenir des résultats exploitables. La première étape de la « décomposition des traits principaux » est la sélection itérative des meilleurs atomes prédicteurs de la dynamique. L'erreur résiduelle de prédiction (*Residual Prediction Error*) d'un résidu $\boldsymbol{i}$ testé en tant que prédicteur est donnée par la variance résiduelle de la matrice $\boldsymbol{\Delta X^{(0)}}$ quand elle est prédite de manière optimale par $\boldsymbol{\Delta X_i^{(0)}}$:

$$\mathrm{RPE}(i) \doteq \min_{A\in\mathcal{M}_{3N\times 3}(\mathbb{R})} \langle \| \Delta X^{(0)} - A\Delta X_i^{(0)} \|^2 \rangle \qquad \textbf{Équation 45}$$

où ⟨ ⟩ est la moyenne empirique calculée sur les trajectoires concaténées de chaque système.

La meilleure matrice de prédiction A_i est donnée par :

$$A_i \doteq \Gamma_{E,i}\Gamma_{i,i}^{-1} \qquad \textbf{Équation 46}$$

avec $\Gamma_{i,i} \doteq \langle \Delta X_i^{(0)} (\Delta X_i^{(0)})^T \rangle, \Gamma_{E,i} \doteq \langle \Delta X^{(0)} (\Delta X_i^{(0)})^T \rangle$.

De là:

$$\mathrm{RPE}(i) = \mathrm{trace}(\Gamma_E - \Gamma_{E,i}\Gamma_{i,i}^{-1}\Gamma_{E,i}^T) \qquad \textbf{Équation 47}$$

où $\Gamma_E = [\Gamma_{E,1} \cdots \Gamma_{E,N}]$ est la concaténation de toutes les matrices $\Gamma_{E,i}$ de taille $3N \times 3$ matrices.

Le meilleur prédicteur $i_* \in argmin_i \mathrm{RPE}(i)$ choisi est celui présentant la plus basse erreur résiduelle de prédiction, donnant le premier prédicteur $i_1 = i_*$. Ensuite, les déplacements atomiques

$$\widehat{\Delta X^{(0)}} \doteq A_{i_*} \Delta X_{i_*}^{(0)} \qquad \textbf{Équation 48}$$

prédits par i_1 sont retranchés de $\Delta X^{(0)}$, ce qui donne la matrice centrée résiduelle

$$\Delta X^{(1)} \leftarrow \Delta X^{(0)} - \widehat{\Delta X^{(0)}}, \qquad \textbf{Équation 49}$$

et la matrice de covariance résiduelle

$$\Gamma_E^{(1)} \leftarrow \Gamma_E - \Gamma_{E,i_*}\Gamma_{i_*,i_*}^{-1}\Gamma_{E,i_*}^T. \qquad \textbf{Équation 50}$$

Cet algorithme de « décomposition des traits principaux » (équations 45 à 48) est répété en remplacant la matrice $\Delta X^{(0)}$ par $\Delta X^{(1)}$, générée par l'équaiton 49. Le second prédicteur i_2 est ainsi déterminé, et la matrice de covariance résiduelle $\Delta X^{(2)}$ est générée. Par itérations successives, la séquence de $\mathcal{P} = \{i_1, \cdots, i_P\}$ est obtenue, jusqu'à atteindre une nombre prédéfini de prédicteur P.

Condensation autour des atomes prédicteurs : La dernière étape de l'approche *PFD* est la condensation des résidus autour du jeu de prédicteurs. Pour chaque atome représentant un résidu (carbone α) $\boldsymbol{i}$, l'algorithme calcule la plus petite erreur de prédiction linéaire normalisée $\boldsymbol{\Delta X_i^{(0)}}$ parmi tous les atomes prédicteurs $\boldsymbol{i_k}$:

$$\mathrm{VPR}(i) \doteq \min_{1 \le k \le P, A \in \mathcal{M}_{3\times 3}(\mathbb{R})} \frac{\langle |\Delta X_i^{(0)} - A\Delta X_{i_k}^{(0)}|^2 \rangle}{\langle |\Delta X_i^{(0)}|^2 \rangle}. \qquad \textbf{Équation 51}$$

Chaque atome est affecté à son meilleur prédicteur i_k si $\mathrm{VPR}(i) \le r$, où r est une valeur seuil résultant en P sous-ensembles $C_1, \cdots, C_P$ qui ne s'intersèctent pas. Une étape de fusion des

sous-ensembles proches est finalement réalisée et un graphe non orienté est généré, où les sous-ensembles C_k sont considérés comme des sommets du graphe. Un lien, noté $C_k \leftrightarrow C_{k'}$, est créé entre les sommets C_k et $C_{k'}$ si $d_{k,k'} \doteq min_{i \in C_k, j \in C_{k'}} \left| \bar{X}_i - \bar{X}_j \right| \leq d_*$. Les sous-ensembles connectés dans le graphe sont regroupés, et fusionnés en un sous-ensemble de plus haut niveau. Ces sous-ensembles de haut-niveau constituent les *IDS*s. Dans le travail présenté, nous avons utilisé les paramètres suivants : $d_* = 5$Å, q = 6 et r = 0.5.

Cette nouvelle méthode de détection des *IDSs* a été utilisée pour la première fois sur STAT5. Les développements en cours concernent la comparaison des résultats obtenus par la décomposition des traits principaux le long de la simulation (entre les dynamiques courtes et les dynamiques longues, entre les dynamiques non concaténées, par rapport à des mouvements stochastiques, *etc.*) et l'étude d'autres systèmes (KIT, VKOR, *etc.*).

V. Recherche de poches

A. Détection de poches et approches analytiques

Dans le contexte moléculaire, une poche désigne un espace dans une protéine ou entre des chaines protéiques dans lequel un ligand/ion/substrat/protéine peut se positionner ou circuler. Il s'agit d'un espace libre et non occupé par les résidus d'une molécule. Approximativement, trois types de poches peuvent être distingués :

- Les canaux peuvent être assimilés à des poches dont le fond est troué dans le sens où ils permettent l'entrée et la sortie du ligand/ion/eau par deux extrémités différentes de la poche. Les canaux sont de manière générale de forme allongée et étroit et peuvent être obstrués facilement par une chaîne latérale d'un résidu, par exemple.
- Les poches de fixation de ligands sont plus larges que les canaux et plus ouvertes, tant qu'aucune molécule n'est fixée. Elles ne possèdent en général qu'une entrée/sortie, et peuvent également être sensibles aux mouvements des chaînes latérales/principales des résidus qui la définissent, voire être transitoire et n'apparaître uniquement dans des conformations particulières.
- Des poches plus larges peuvent exister, généralement à la surface des macromolécules. Ces poches sont en générale ouvertes et correspondent à des sites non spécifiques.

La détection des poches est un enjeu important de la biologie, en particulier dans le cadre des recherches visant à inhiber des protéines impliquées dans les processus physiopathologiques où elles constituent un élément crucial à la conception de nouvelles molécules inhibitrices. De même, l'évaluation des différentes propriétés (volume, résidus à la marge, positionnement dans une protéine,…) de cette poche est importante car de ces paramètres vont découler les fonctionnalités nécessaires pour accueillir un ligand. Ces données sont très importantes dans les projets de caractérisation des pharmacophores ou dans les études de relation structure-activité quantitatives.

Plusieurs algorithmes ont été décrits afin de suivre les propriétés des poches détectées à la surface des protéines et/ou nucléotides et l'évolution de leur forme et volume, qui se basent sur des critères soit géométriques[579–588], soit énergétiques[589,590]. Nous détaillons ci-dessous la base méthodologique principale du logiciel Fpocket et de son dérivé MDpocket, qui présentent l'avantage de mesurer plusieurs paramètres relatifs à la poche (volume, surface accessible au solvant,…) mais également d'extraire un ensemble de données pour chaque conformation issue de simulation de dynamique moléculaire (liste des résidus définissant la poche, scores, …).

B. Fpocket et MDpocket en détail

Les poches des STAT5s monomériques (STAT5a, p-STAT5a, STAT5b et p-STAT5b) ont été détectées avec l'outil MDpocket [591,592], qui se base sur l'algorithme de Fpocket[585]. Cette algorithme applique la théorie des sphères α [582] sur une grille superposée au système. Pour chaque conformation issue des trajectoires de dynamique moléculaire, Fpocket associe une sphère α (une sphère contactant au moins quatre atomes sans contenir aucun autre atome à l'intérieur) au point de la grille le plus proche. La taille des sphères peut être contrôlée afin de détecter des poches de taille différente. Les sphères α ainsi placées sont ensuite regroupées selon des critères de distance et de nombre afin de former des poches qui correspondent aux critères de recherche. En fonction des différents critères utilisés pour le regroupement des sphères α, différents types de poches sont détectées (canaux, poches au sein d'une protéine, poches à la surface), et les descripteurs associés à ces poches sont ensuite sauvegardés pour analyse. Enfin, les cartes de densité et de fréquence relatives aux poches sont produites par itération de cette étape sur l'ensemble des conformations des trajectoires. La carte de densité représente la densité en sphère α, alors que la carte de fréquence est liée à la fréquence à laquelle chaque point de la grille est occupé par une sphère α. Afin de limiter l'apparition d'artefacts liés aux déplacements de la protéine, chaque conformation a été superposée à une structure de référence (conformation moyenne).

Pour analyser la poche de liaison proche de résidu phosphotyrosyl, des carbones α du domaine SH2 ont été utilisés, notamment les résidus K600, R618 – E623, N642 and K/M644 (pour STAT5a et STAT5b, respectivement). Les poches sélectionnées ont été ultérieurement analysées au cours d'un second calcul. Pour chaque poche analysée, les points de la grille associés plus de 25 % du temps de simulation à une ou plusieurs sphères α ont été déclarés pour l'analyse du volume de la poche.

Les scores de conservation des résidus de STAT5 comparé aux autres protéines STATs humaines ont été calculés par le serveur web ConSurf[593–595] ; les séquences primaires des STATs humaines pour ces calculs ont été obtenues à partir de la base de données des protéines du NCBI (*National Center for Biotechnology Information*, http://www.ncbi.nlm.nih.gov/protein).

Chapitre 3 : Résultats

I. Dynamique moléculaire intrinsèque des monomères de STAT5

A. Variations de structure et dynamique de STAT5

1. Analyse des structures des protéines de la famille STAT

Après des recherches extensives au sein de la base de données structurale PDB, il nous est apparu qu'aucune structure n'était disponible pour les protéines humaines STAT5. Dix structures recouvrant partiellement la séquence des protéines STAT5 humaines ont néanmoins été extraites de la *Protein Data Bank*[441] (*cf.* paragraphe II.B.2 du chapitre 1 et Tableau 3), rapportant des protéines de la famille des STATs humaines (STAT1, STAT2 et STAT6) ou murines (STAT3, STAT4 et STAT5). L'analyse des structures disponibles au sein de la PDB a révélé que tous les domaines des protéines STATs ne sont pas caractérisés.

Le domaine N-terminal, long d'environ 130 à 135 résidus chez les STATs, est caractérisé dans deux structures cristallographiques. Le domaine N-terminal de STAT4 murin a été caractérisé sous forme dimérique (code PDB : 1BGF)[138], et révèle une structure composée de 8 hélices organisées en forme de crochet (*cf.* Figure 22). Les chaines de chaque monomère interagissent fortement *via* des contacts polaires à la surface d'une des faces du « crochet », qui permettent la polymérisation de dimères liés à des sites adjacents [143]. La mutation d'un résidu de cette interface réduit considérablement la transcription *in vitro* des gènes en réponse à une stimulation par l'interféron-γ [138], démontrant le rôle joué par ce domaine. La seconde structure cristallographique caractérisant le domaine N-terminal est la forme humaine de STAT1 (code PDB : 1YVL)[117], qui contient plusieurs domaines structuraux autres que le domaine N-terminal, à l'inverse de la structure 1BGF. Les autres domaines résolus sont le CCD, DBD, LD et le domaine SH2, soit les domaines correspondant au *Core Fragment* (*CF*) des protéines STATs. Les résidus de la queue tyrosyl – le système n'est pas phosphorylé – (les résidus 684 à 712 de STAT1 correspondent aux résidus 685 à 711 de STAT5), bien que présents dans les constructions, n'ont pu être résolus, indiquant une forte flexibilité de cette région. L'ensemble de la structure révèle un arrangement tétramérique, bien que non lié à l'ADN, consistant en un assemblage de deux dimères anti-parallèles. Chaque dimère adopte une configuration en « bateau », dont les interfaces sont constituées par les domaines CCD et DBD, le CCD du monomère A interagissant avec le DBD du monomère B et *vice versa*. Une seconde interface, décrite entre les domaines N-terminaux, présente une grande similarité avec l'interface ND – ND observée dans la structure 1BGF. Cependant, la boucle entre le domaine N-terminal et le *Core Fragment* n'a pas été résolue. De plus, cette espèce tétramérique, au sein de laquelle des interactions ND – ND sont clairement définies, semble contraindre la position du ND dans une conformation qui n'est pas forcément

celle qu'il adopterait *in vivo* dans un arrangement monomérique. Pour cette raison, nous n'avons pas inclus le domaine N-terminal dans nos modèles générés par homologie.

Le *Core Fragment* a également été résolu dans la forme phosphorylée de STAT1 humain représentant le dimère parallèle lié à l'ADN (code PDB : 1BF5)[133], dans la forme phosphorylée, dimérique parallèle et liée à l'ADN de STAT3 murin (code PDB : 1BG1)[132], dans la forme non-phosphorylée de STAT3 murin, stabilisée en dimère antiparallèle sans ADN (code PDB : 3CWG)[121], dans la forme non-phosphorylée de STAT3 murin, formant un dimère parallèle lié à l'ADN (code PDB : 4E68)[134], et dans la forme non-phosphorylée de STAT5a murin observée comme dimère antiparallèle sans ADN (*cf.* Figure 3), code PDB : 1Y1U)[118]. Les domaines du *CF* présentent une très forte similarité structurale entre les différentes protéines, en dépit de leur divergence au cours de l'évolution.

Le domaine CCD présente ainsi quatre hélices α de grande taille (*cf.* Figure 31), dont les résidus tournés vers l'extérieur sont majoritairement hydrophiles et les résidus tournés vers les autres hélices sont principalement hydrophobes. Une différence notable cependant est la longueur des deux premières hélices, *α1* et *α2*, qui sont sensiblement plus longues dans la structure1Y1U comparativement aux autres structures du *CF*, et qui présentent une courbure. L'alignement des séquences primaires de STAT humaines indique une insertion de 12 à 20 résidus entre les deux hélices *α1* et *α2* chez les deux isoformes de STAT5 humains par rapport aux autres protéines STATs, ce qui pourrait se traduire par une augmentation de la longueur de la boucle *α1 – α2* et/ou par une augmentation de la longueur des deux hélices. La très forte identité de séquence de cette région (résidus 185 à 205) avec la forme murine de STAT5a (90 % par rapport à STAT5a humain, et 80 % comparé à la forme humaine de STAT5b) nous incite à penser que la structure tridimensionnelle des STAT5s humaines est fortement similaire à celle de la forme murine de STAT5a.

Le domaine de liaison à l'ADN (DBD) montre une structure composée d'environ 10 feuillets β dans toutes les structures connues, avec l'insertion de petites hélices 3_{10} dans les boucles reliant les feuillets (*cf.* Figure 31). Ces boucles reliant les divers éléments structuraux peuvent être longues (jusqu'à 20 résidus), et ne sont pas toutes résolues entièrement dans toutes les structures. Lorsqu'elles le sont, les facteurs de température associés à ces résidus sont les plus élevés de cette région, voir par exemple la structure (1Y1U, 4E68).

Le domaine de liaison (LD) est composé principalement d'hélices α (*cf.* Figure 31), entre lesquelles des boucles de taille moyenne (<15 résidus) sont trouvées. Un court feuillet β est également conservé dans toutes les structures cristallographiques à l'exception de la structure 3CWG.

Le domaine SH2 (SRC-Homology 2) est un domaine relativement générique, présent dans plus de 100 protéines liées à la régulation des voies de signalisation faisant appel à la phosphorylation de résidus tyrosine[596,597]. Ils ont un rôle de reconnaissance des résidus phosphotyrosine dans les cellules de métazoaires, en parallèle avec d'autres modules comme, par exemple, les domaines C2 [598], le module pyruvate kinase M2[599]. Les domaines SH2 des STATs

présentent un repliement mixte de feuillets β et d'hélices α et/ou 3_{10} de taille limitée (<15 résidus) dans les structures cristallographiques de la PDB (*cf.* Figure 31).

L'ensemble de ces structures de STATs présentent un point commun, à savoir l'absence de la boucle C-terminale du domaine SH2, correspondant au début de la queue (phospho)-tyrosyl. Dans certains cas, cette boucle est tronquée lors de la construction du système à cristalliser, alors que dans d'autres cas (dimères parallèles liés à l'ADN de STAT3 ou STAT1, et dimère antiparallèle de STAT5a murin), elle est présente dans les constructions initiales, mais non-résolue. Ces résultats orientent vers une très forte variabilité de la position des atomes au niveau de cette boucle.

Finalement, le TAD n'est caractérisé que dans trois structures résolues par spectroscopie de Résonance Magnétique Nucléaire (RMN). Ces structures sont partielles, ne couvrant que les résidus 706 à 751 de STAT1 murin (code PDB : 2KA6), les résidus 782 à 838 de STAT2 murin (code PDB : 2KA4)[526], les résidus de 795 à 808 de STAT6 murin (code PDB : 1OJ5)[525]. De plus, ces structures sont résolues en complexe avec divers partenaires protéiques (CBP/p300 ou NCoA-1), dont nous ne souhaitons pas étudier les effets sur la structure de STAT5. Enfin, sachant que cette région de la protéine est la plus variable en termes de séquence protéique, nous n'avons pas introduit ce domaine dans les modèles générés par homologie.

Si les domaines N- et C-terminaux de STAT5 n'ont pas été pris en considération, nous avons généré des modèles par homologie du *Core Fragment* en utilisant (i) la structure cristallographique la plus proche en termes d'identité de séquences (1Y1U) avec les isoformes humaines de STAT5a et STAT5b, (ii) la structure la plus finement résolue (*i.e.* celle avec la plus haute résolution) et pour laquelle l'environnement de la tyrosine critique (le site de phosphorylation) est le mieux caractérisé (1BG1).

2. Analyse des structures des modèles de STAT5 générés par homologie

Quatre modèles de STAT5 ont été générés : STAT5a non-phosphorylé (STAT5a), STAT5a phosphorylé (pSTAT5a), STAT5b non-phosphorylé (STAT5b) et STAT5b phosphorylé (pSTAT5b). Etant donné la forte identité des séquences entre les deux isoformes STAT5a et STAT5b (>93%), les modèles générés présentent une forte similarité structurale, caractérisée par la forme générale de la protéine, où 3 domaines sont situés dans un même axe (CCD, LD et SH2), alors que le DBD et la queue (phospho-)tyrosyl sont à la marge, en décalage de de cet axe (*cf.* Figure 31).

Le domaine CCD est composé de 4 hélices (*α1- α4*) dans nos modèles (*cf.* Figure 31). *α1* (46 ou 50 résidus) et *α2* (65 à 69 résidus) présentent une longueur légèrement supérieure (38 et 64 résidus pour *α1* et *α2*, respectivement) et une courbure similaire à la structure 1Y1U. Les hélices *α3* et *α4* sont de taille similaire aux structures supports, entre 28 et 30 résidus et entre 21 et 23 résidus, respectivement, et ne présentent pas de courbure significative. Le domaine CCD contacte le DBD par les hélices *α3* et *α4*, qui touchent les boucles *bc* et *fg* pour former un cœur

hydrophobe. Le CCD contacte également le LD, avec l'extrémité C-terminale de l'hélice *α2* et la boucle suivante qui touche les hélices *α5* et *α6*. Le domaine de liaison à l'ADN est composé de 8 brins β constants constituants 3 feuillets : les brins *a* (résidus 333-337), *b* (résidus 349-355) et *e* (résidus 413-420) forment le feuillet *abe*, les brins *c* (résidus 369-375), *f* (résidus 442-448) et *g* (résidus 456-463) forment le feuillet *cfg*, tandis que les brins *a'* (résidus 342-344) et *g'* (résidus 468-470) forment le dernier feuillet, qui contacte le domaine LD (hélice α5). Les boucles reliant les brins *c* et *d*, et les brins *e* et *f* sont les plus longues et sont orientées vers l'extérieur de la protéine. D'autres feuillets composés de brins très courts ont été détectés (*cf.* Figure 32). Le LD est invariablement composé de 7 hélices (dénommées *α5*, *α6*, *α7*, *α7'*, *α7''*, *α8* et *α9*) et d'un feuillet de brins antiparallèles (i et h). Le domaine SH2 repose sur le plan formé par les trois hélices *α7*, *α8* et *α9*. Le domaine SH2 est assez variable, mais présente un feuillet unique composé de 2 ou 3 brins (*A*, *B* et *C*) entouré de 3 à 5 hélices α ou 3_{10}. Les modèles non-phosphorylés présentent un feuillet à 2 brins alors que les deux modèles phosphorylés présentent un feuillet à 3 brins. Cette variation correspond à une position différente (~1,4 Å) des résidus L643 et K/M644 (brin *C*), alors que les résidus T628 et I629 (brin *B*) sont presque superposés dans tous les modèles (< 0,6 Å). Enfin, la queue (phospho)-tyrosyl (résidu 694 ou 699 chez STAT5a et STAT5b, respectivement) ne présente aucun repliement régulier, de type hélice ou feuillet. Les résidus 694 de STAT5a et 699 de STAT5b sont spatialement proches lorsque les modèles sont superposés, alors que les résidus phosphotyrosine sont éloignés l'un de l'autre, et éloignés également de la position des résidus tyrosine correspondant.

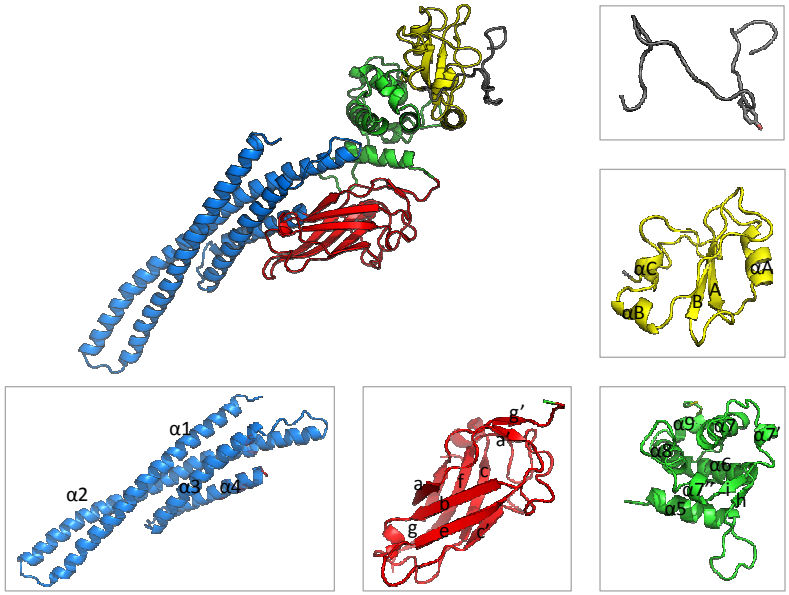


***Figure 31:* Modèle de STAT5b généré par homologie.** Le CCD est en bleu, le DBD est en rouge (), le LD est vert, le SH2 est en jaune et la queue tyrosyl est en gris. La chaîne latérale de la tyrosine 699 est tracée en bâton dans la fenêtre.

Les deux isoformes du *CF* de STAT5 diffèrent par 16 résidus dans les différents domaines et par une insertion de 5 résidus (CESAT) dans la queue (phospho)-tyrosyl. Les 16 remplacements sont situés dans la boucle *α1-α2* (A187G-Q188P), dans la boucle entre les brins *c* et *c'* (résidus E391D-C392Y), *e* et *f* (A427S), dans le brin *f* (V442I), dans la boucle reliant les brins *f* et *g* (S452G), dans l'hélice *α5* (H476N), dans la boucle entre les hélices *α7''* et *α8* (W566R), à

l'extrémité C-terminale de l'hélice *α8* (H585L), dans la boucle entre les brins *B* et *C* du domaine SH2 (P636Q, N639M-L640F) ou dans le brin *C* (K644M), entre l'hélice *αC* et *αD* (S664N) ou à l'extrémité de l'hélice *αD* (F679Y). Malgré ces différences de séquence, les modèles générés ne diffèrent que faiblement, caractérisés par un RMSD sur les carbones α variant de 0,61 à 1,34 Å sur l'ensemble de la structure. Les déviations par rapport aux structures supports sont du même ordre, même si les modèles générés sont plus proches de la structure 1Y1U (RMSD compris entre 0,309 et 1,259Å, *cf.* Tableau 6) que de la structure 1BG1 (RMSD dans l'intervalle 1,855 et 2,101, *cf.* Tableau 6).

Tableau 6: **Distances des modèles aux structures supports.** Les distances sont exprimées en angströms (Å). Les distances sont calculées en prenant les carbones α de tous les résidus, ou en excluant les résidus de la queue (phospho-)tyrosyl (entre parenthèses).

	STAT5a	pSTAT5a	STAT5b	pSTAT5b
1Y1U	0,90 (0,31)	1,58 (1,08)	1,93 (0,54)	2,57 (1,26)
1BG1	3,08 (1,86)	3,09 (2,03)	3,08 (2,10)	3,10 (2,08)

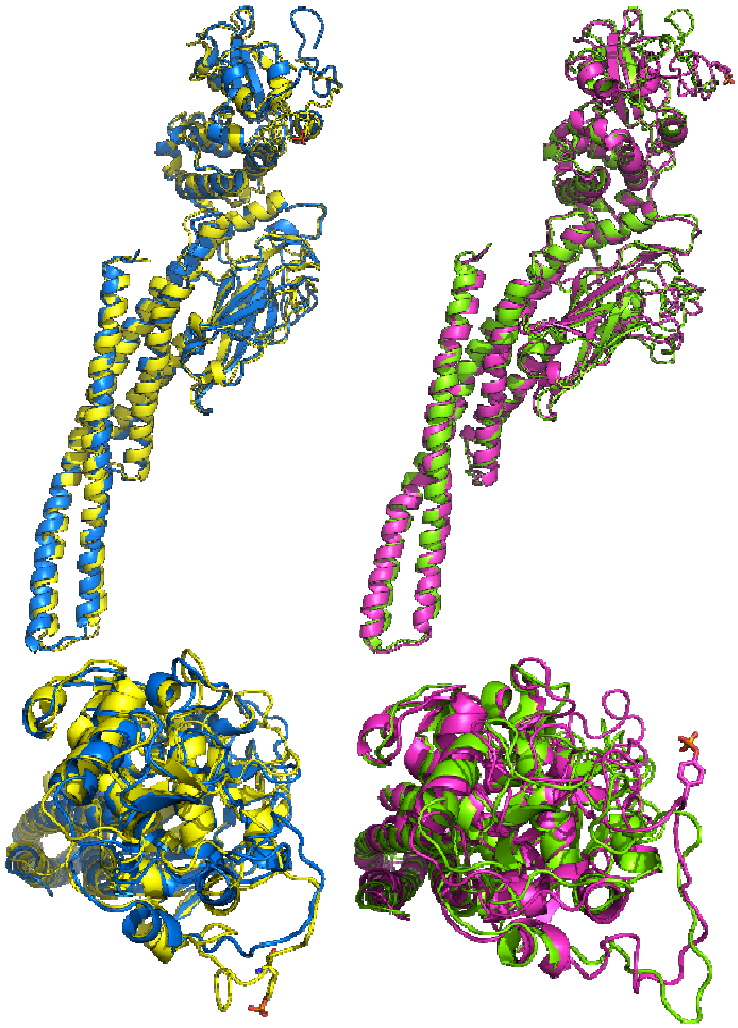


Figure 32: **Superposition des modèles monomériques générés par homologie.** STAT5a (en bleu) et pSTAT5a (en jaune) ; , STAT5b (en vert) et pSTAT5b (en magenta) . Les protéines sont représentées en 'cartoon' ; les chaînes latérales des résidus de phosphotyrosine sont représenté en bâtons.

3. Stabilité des dynamiques de production:

La dynamique moléculaire (DM) de chaque modèle généré a été simulée. L'ensemble du temps de calcul obtenu auprès des organismes nationaux a été consommé, et complété auprès de plusieurs partenaires (société BULL, université de Rio de Janeiro). Les deux simulations de DM de production de chaque protéine ont été analysées à la fois en termes de comportement dynamique globale (RMSD par rapport à la structure initiale et à la structure moyenne), interne (RMSD par domaine) ou locale (RMSF).

Les profils des simulations de DM ont d'abord été analysés de manière globale : la superposition des conformations de dynamique moléculaire sur la structure de référence a été réalisée en utilisant l'ensemble des carbones α de la protéine. La structure de référence utilisée est soit la structure initiale des simulations (à t = 0 ns, *cf.* Figure 33), soit la structure moyenne de la simulation (*cf.* Figure 34). Les déviations (RMSDs) observées par rapport aux structures initiales augmentent rapidement pour atteindre un plateau situé à environ 4 Å autour duquel elles oscillent de manière plus ou moins ample (*cf.* Figure 33, courbes noires). Ce comportement semble indiquer une convergence globale au cours de la dynamique mais indiquent aussi que d'autres effets ont lieu, ce qui explique la variation de RMSDs autour du plateau. Les valeurs de RMSD calculées en prenant en référence la structure moyenne montrent un plateau constant (*cf.* Figure 34, courbes noires), mais fluctuant sensiblement à la manière des valeurs calculées comparativement à la structure initiale (*cf.* Figure 33, courbes noires). Ces résultats corrèlent bien avec des déviations de certains domaines au sein de la protéine, dans laquelle des régions présentent des variations importantes sans entrainer de réarrangement majeur. Les RMSDs calculés par domaine en utilisant un alignement global (en utilisant l'ensemble de la protéine) montrent clairement que les déviations ne sont pas réparties de manière uniforme mais principalement reliées à la queue porteuse du résidu (phospho-)tyrosyl (*cf.* Figure 33 et Figure 34, courbes grises), et dans une moindre mesure par le CCD (*cf.* Figure 33 et Figure 34, courbes bleues).

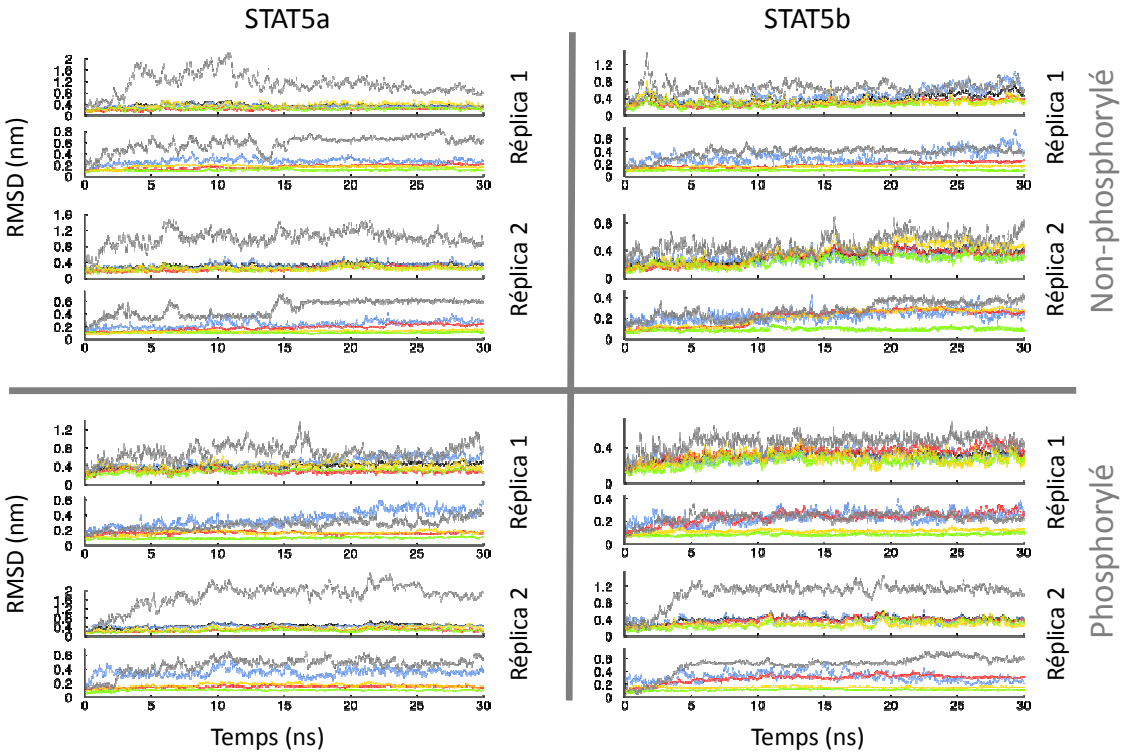


Figure 33: **Déviations au cours des simulations de dynamique moléculaire des protéines STAT5 par rapport à la conformation initiale (t = 0 ns).** Les deux répliques de 30 ns de STAT5a sont dans le cadran supérieur gauche, pSTAT5a dans le cadran inférieur gauche, STAT5b dans le cadran supérieur droit et pSTAT5b dans le cadran inférieur droit. Pour chaque réplica, deux séries de courbes sont présentées : en haut, l'étape de superposition est faite en utilisant l'ensemble des carbones α puis le RMSD est calculé pour l'ensemble de la protéine (courbe noire) et pour chaque domaine (CCD est en bleu, DBD est en rouge, LD est en vert, SH2 est en jaune, queue (phospho-)tyrosyl est en gris). En bas, la superposition est réalisée pour chaque domaine avant calcul du RMSD pour le même domaine (le code couleur est identique). Les échelles de RMSDs peuvent varier d'une figure à l'autre.

La queue présente dans tous les STAT5s non-phosphorylés une grande variabilité due à la fois à l'absence de structures secondaires qui pourraient rigidifier cette région et ainsi limiter les fluctuations atomiques observées, à la forte exposition au solvant permettant une forte mobilité et à son extrémité C-terminale libre. On peut penser que la présence du domaine de transactivation (TAD), long de plus de 80 résidus, modifie de manière non–négligeable la dynamique de cette région. La flexibilité de cette boucle et l'absence de structures secondaires sont néanmoins compatibles avec le rôle fonctionnel de cette portion, à savoir permettre au résidu de phosphotyrosine de se fixer sur un monomère et ainsi de favoriser la dimérisation parallèle. Les valeurs de RMSD observées dans le cas d'une superposition globale (en utilisant l'ensemble des carbones α du système) sont supérieures (de 0,4 à 2,2 nm, *cf.* Figure 33, graphiques du haut pour chaque réplica, courbes grises) à celles que nous observons lorsque nous utilisons uniquement les atomes de carbone α de la queue (de 0,2 à 0,8 nm, *cf.* Figure 33, graphiques du bas pour chaque réplica, courbes grises). Cette observation indique que l'extrémité C-terminale de STAT5 présente parfois une mobilité importante en début de simulation de DM avant de trouver un repliement métastable, qu'il adopte pendant le reste de la simulation. La conformation moyenne de la protéine pour cette région sera de fait proche de

l'état métastable, expliquant les RMSDs inférieurs lorsque la structure de référence est la structure moyenne.

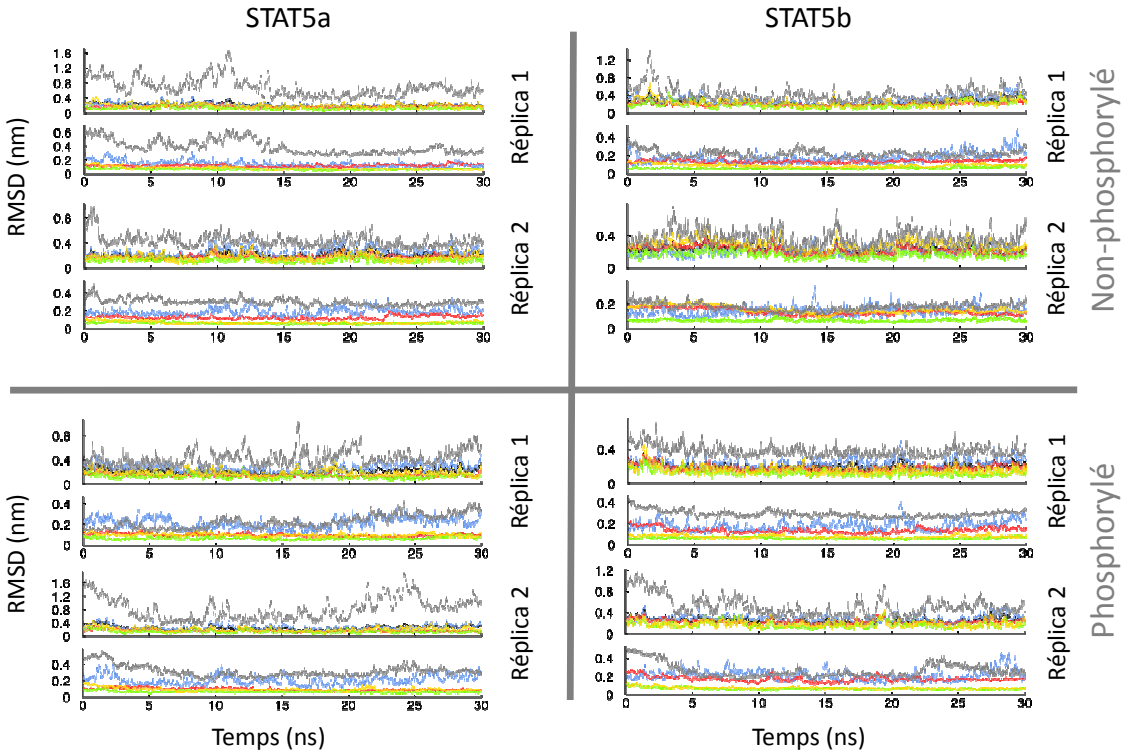


Figure 34: **Déviations au cours des simulations de dynamique moléculaire par rapport à la conformation moyenne.** Les deux répliques de 30 ns de STAT5a sont représentées dans le cadran supérieur gauche, pSTAT5a dans le cadran inférieur gauche, STAT5b dans le cadran supérieur droit et pSTAT5b dans le cadran inférieur droit. Pour chaque réplica, deux séries de courbes sont présentées : en haut, l'étape de superposition est faite en utilisant l'ensemble des carbones α puis le RMSD est calculé pour l'ensemble de la protéine (courbe noire) et pour chaque domaine (CCD est en bleu, DBD est en rouge, LD est en vert, SH2 est en jaune, queue (phospho-)tyrosyl est en gris). En bas, la superposition est réalisée pour chaque domaine avant calcul du RMSD pour le même domaine (le code couleur est identique). Les échelles de RMSDs peuvent varier d'une figure à l'autre.

Le domaine CCD montre une certaine flexibilité, moins marquée cependant que la queue C-terminale (*cf.* Figure 33 et Figure 34, courbes bleues). La présence d'hélices α dans ce domaine explique les déviations moins marquées de RMSD. Cependant, la présence de déviations importantes au sein d'un domaine aussi structuré dénote la présence d'effets dynamiques notables. Les valeurs de RMSD utilisant une superposition globale (*cf.* Figure 33 et Figure 34, graphiques du haut pour chaque réplica, courbes bleues) sont sensiblement plus élevées que celles obtenues lors d'une superposition locale (*cf.* Figure 33 et Figure 34, graphiques du bas pour chaque réplica, courbes bleues), ce qui montre que le CCD présente des mouvements notables par rapport au reste de la protéine mais une dynamique locale limitée. La visualisation des dynamiques a montré plus précisément que la région distale du CCD, composée des extrémités C-terminales de l'hélice *α1* et N-terminale de l'hélice *α2* ainsi que de la boucle reliant ces deux hélices et correspondant aux résidus 180 à 220, présente un mouvement d'oscillation considérable par rapport au reste de la protéine. Ce mouvement, parfois ample, explique les variations de RMSD et leur ampleur.

Les trois autres domaines, DBD, LD et SH2, présentent de manière générale des variations de RMSD beaucoup plus faibles, quel que soit le type de superposition utilisé, dénotant une grande stabilité au cours du temps de simulation. Ces domaines sont situés autour du centre géométrique de la protéine, ce qui fait que les mouvements vont affecter de manière moins marquée les valeurs de RMSD globaux (calculés en utilisant l'ensemble des carbones α) liées aux fluctuations des extrémités du système. Dans le cas des RMSDs locaux (calculés en superposant chaque domaine de manière individuelle), ces domaines montrent des valeurs stables de RMSD inférieures à 2 Å (*cf.* Figure 33 et Figure 34, courbes rouge, verte et jaune).

Pour toutes ces simulations de dynamique moléculaire, les réplicas se comportent de manière similaire, montrant une excellente reproductibilité. De manière similaire, les RMSDs calculés pour chaque domaine sont proches, indiquant une dynamique interne stable indépendamment de l'isoforme de STAT5 considérée ou du statut de la tyrosine (phosphorylée ou non). Les éléments spécifiques de chaque protéine STAT5 (remplacements/variations, insertion d'une séquence courte ou ajout du groupement phosphate) ne semblent pas apporter de modifications notables de la dynamique interne.

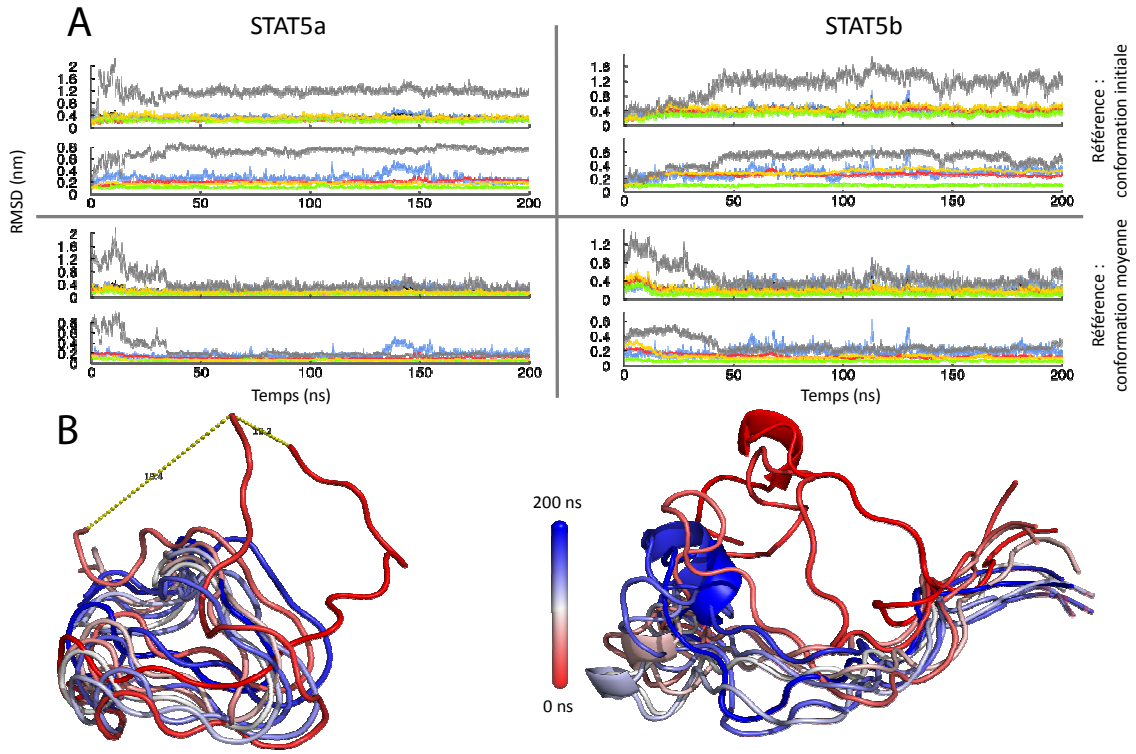


Figure 35: **Déviations (RMSDs) au cours des simulations de DM prolongées, de 200ns.** (A) Chaque cadran présente deux séries de courbes : en haut, les RMSDs sont calculés en utilisant une superposition globale (en utilisant l'ensemble des carbones α) ; en bas, les RMSDs sont calculés en utilisant une superposition de chaque domaine. (B) Pour chaque trajectoire, 11 conformations espacées de 20 ns (de t = 0 ns en rouge , à t = 200 ns en bleu) ont été superposées afin de montrer l'évolution de la position de la queue tyrosyl au cours de la simulation de dynamique moléculaire.

Les simulations de dynamique moléculaire étendues à 200 ns ont révélé un comportement similaire aux simulations de 30 ns des protéines, que ce soit au niveau global ou au niveau des domaines étudiés individuellement (*cf.* Figure 35A). À l'exception de la queue tyrosyl, les deux protéines s'équilibrent rapidement (après 10 ns) autour d'un plateau de 3,6 Å (STAT5a) ou de 4,5 Å (STAT5b) comparativement à la conformation initiale des simulations, et d'un plateau de 1,8 Å (STAT5a) ou 2,4 Å (STAT5b) par rapport à la structure moyenne des simulations (*cf.* Figure 35A, courbes noires). Chaque domaine de STAT5 présente également des variations internes, le CCD et la queue tyrosyl présentant toujours les plus grands déplacements (*cf.* Figure 35A, courbes bleues et grises, respectivement). Ce dernier domaine présente un temps d'équilibration plus long que le reste de la protéine, de l'ordre de 40 à 50 ns (*cf.* Figure 35A, courbes grises). La queue tyrosyl va en effet se déplacer pour adopter un repliement favorable énergétiquement, qui se traduit par une stabilisation du RMSD à partir de 40 ns (STAT5a) ou 50 ns (STAT5b). Le CCD présente des valeurs de RMSDs globalement stables, mais des pics de RMSDs peuvent être observés (entre 140 et 150 ns pour STAT5a, à 112 et 128 ns pour STAT5b, *cf.* Figure 35A), ce qui correspond à des déplacements extrêmes du CCD distal par rapport au CCD proximal (et au reste du système) constituant le CCD.

La flexibilité de chacun des résidus a été évaluée par la mesure du RMSF des carbones α (*cf.* Figure 36). Bien que ne prenant pas en compte les fluctuations atomiques des chaînes latérales des résidus, cette mesure est une approche crédible pour repérer les résidus les plus mobiles. La prise en compte des chaînes latérales peut introduire de fortes fluctuations, notamment dans le cas des résidus à longue chaîne latérale possédant des liaisons pouvant pivoter (arginine, lysine, méthionine, glutamine…). Mesurer les fluctuations des carbones α donne ainsi une indication claire et non bruitée de la mobilité de la chaîne principale d'une protéine. Les différences de mobilité très importantes entre les régions les moins fluctuantes (hélices α du CCD, LD, feuillet β du domaine SH2), les régions présentant des fluctuations moyennes (boucles entre les brins des feuillets du DBD, boucles du domaine SH2) et les régions extrêmement fluctuantes (extrémité N-terminale du *Core Fragment*, CCD distal et queue – terminale) sont en accord avec les observations issues de celles des courbes de RMSD.

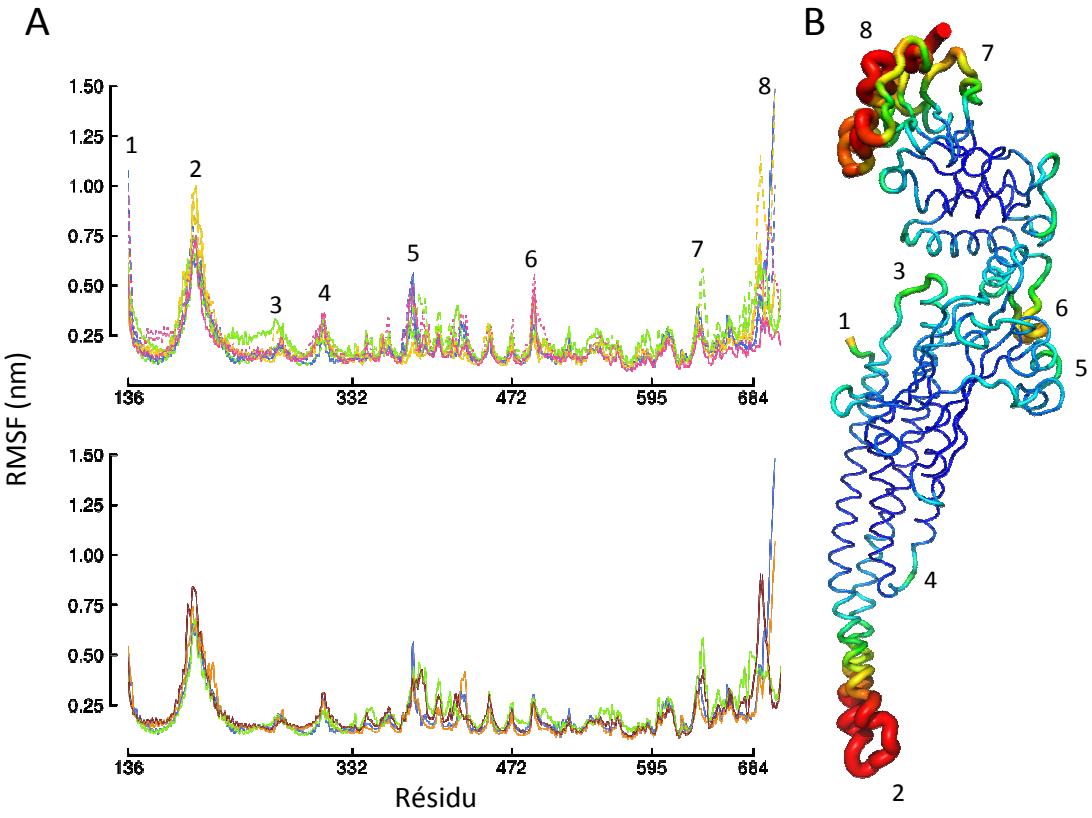


Figure 36: **Fluctuations (RMSFS) atomiques des STAT5.** Les RMSFs ont été calculés pour chaque carbone α de la protéine, et projection sur la structure de STAT5a. (A - haut) RMSFs des simulations de 30 ns. Les répliques de STAT5a sont en bleu, pSTAT5a en jaune, STAT5b en vert et pSTAT5b en magenta. Pour chaque système, les deux répliques sont tracées en trait plein et en pointillés. (A – bas) RMSFs des simulations de 200 ns. STAT5a est affiché en orange et STAT5b en marron. Le RMSF des trajectoires de 30 ns sont reproduites. (B) Structure de STAT5a, la largeur du ruban est proportionnelle à la valeur du RMSF. Les chiffres assurent la correspondance entre les pics observés en (A) et leur position spatiale (B).

L'extrémité N-terminale présente des grandes valeurs de RMSFs, qui s'explique par l'absence de structure secondaire et la présence d'une extrémité libre (*cf.* Figure 36A et B, pic 1). La formation de l'hélice α1 engendre une structuration importante des résidus et une fluctuation beaucoup plus faible des résidus proches, ce qui explique la diminution très rapide de ce pic de RMSF. Un second pic qui présente des valeurs de RMSF élevées, comprises entre 6 et 10 Å, est constamment observé à hauteur du CCD distal (*cf.* Figure 36, pic 2). Comparé aux autres résidus des hélices α1 et α2 dont les fluctuations sont à des niveaux faibles (~2 Å), ces données permettent de dire que le CCD est un domaine dynamique composé de deux modules dont l'un est mobile tandis que l'autre est beaucoup plus stable. Le CCD distal constitue la partie mobile qui oscille autour de la partie fixe, le CCD proximal constitué de la partie N-terminale de l'hélice α1, de la partie C-terminale de l'hélice α2 et du reste du CCD. Cette dualité du comportement dynamique explique les variations de RMSDs observé, à l'échelle locale (dans le CCD) mais aussi en partie à l'échelle globale (dans le CF), sur toutes les simulations étudiées. Le CCD distal correspond à des résidus présentant de fortes valeurs de facteur de température dans le cristal 1Y1U, ce qui corrèle bien avec nos résultats et valide la justesse de nos calculs. Enfin, le dernier pic (*cf.* Figure 36, pic 8) est le plus important, recouvre la queue (phospho-)tyrosyl, dont

nous savons qu'elle présente une forte mobilité par rapport aux structures initiales ou moyennes. Toutes les simulations ne présentent pas le même profil dans cette région : certaines simulations indiquent une queue (phospho-)tyrosyl très mobile (RMSFs supérieurs à 10 Å), tandis que d'autres présentent des RMSFs moyens, inférieurs à 4 Å. Les plus hautes valeurs de RMSFs pour cette région sont observées pour le premier réplica de STAT5a (*cf.* Figure 36A – haut, courbe bleue) dont on peut voir qu'elle subit un fort déplacement durant les premières 40 ns (*cf.* Figure 35B – gauche). Le RMSF élevé reflète ainsi cette dynamique, alors que dans d'autres protéines, la queue trouve un état dynamique stable avec une amplitude de déplacement équivalente, ce qui limite les déplacements au cours de la simulation réalisée.

Les pics 3 et 4 correspondent à deux autres boucles du CCD qui relient les hélices *α2* et *α3*, et les hélices *α3* à *α4*, respectivement. Le pic 5 correspond à la longue boucle du domaine de liaison à l'ADN qui relie les brins *c* et *c'*, et est fortement exposée au solvant, ce qui engendre une plus grande mobilité, comparativement au reste du domaine. Le pic 6 correspond à la longue boucle reliant l'hélice *α5* au brin h, qui est également accessible au solvant. Enfin, le pic 7 correspond à la boucle reliant les brins *B* et *C* du feuillet β du domaine SH2 (*cf.* Figure 36).

Ces résultats corrèlent avec les facteurs de température des structures supports 1Y1U et 1BG1. Dans ces structures, les plus grandes fluctuations sont observées pour deux boucles du domaine de liaison à l'ADN (reliant les brins *c* et *c'*, et les brins *e* et *f* dans la structure 1BG1, cette boucle n'étant pas résolue dans la structure 1Y1U), la partie distale du CCD (dans la structure support 1Y1U), la boucle reliant les brins *B* et *C* du domaine SH2, ainsi que la partie C-terminale de la queue phospho-tyrosyl (présente uniquement dans la structure 1BG1). Ces données corrèlent bien aux données générées par DM, puisque les régions présentant les fluctuations les plus élevées sont superposables.

L'ensemble des trajectoires de DM présentent un profil commun et similaire, indépendamment du statut de phosphorylation des protéines ou de l'isoforme de STAT5 considéré. Les dynamiques des trajectoires longues (de 200 ns) présentent également le même profil que celui des trajectoires courtes (30 ns). Les variations de RMSDs et de RMSFs sont différentes lorsque l'on compare une dynamique courte et son extension jusqu'à 200 ns au niveau de l'extrémité C-terminale. Ces différences reflètent la stabilisation progressive de cette région au cours de la dynamique étendue. Finalement, toutes les protéines STAT5 possèdent des caractéristiques dynamiques communes qui se traduisent par :

- une queue (phospho-)tyrosyl très mobile qui visite diverses positions, explorant un grand espace conformationnel,
- trois domaines – SH2, LD et DBD – qui sont très stables au cours des simulations de DM mais dont les boucles les plus longues et les plus exposées au solvant sont plus flexibles,
- un domaine présentant deux pseudo-modules dont les dynamiques diffèrent en dépit de la présence de structures secondaires stables.

4. *Impact du groupement Phosphate et différences dépendantes de la séquence*

La dynamique générale des deux isoformes est très similaire, à l'échelle de la protéine (*CF*) ou des domaines séparés. Cependant, d'autres manifestations de la dynamique moléculaire peuvent être observées et quantifiées *via* l'évolution des structures secondaires de la protéine au cours des simulations. Si le repliement global de la protéine reste inchangé par rapport aux modèles générés par homologie, des différences plus subtiles peuvent être notées (*cf.* Figure 37). Le logiciel DSSP[560,561] différencie les feuillets β des ponts β en fonction du nombre de liaisons hydrogènes : si plus de trois liaisons hydrogènes consécutives sont observées entre deux brins, les résidus forment un feuillet β, alors que les segments plus courts seront considérés comme des ponts β.

Différences observées entre réplicas :

Chaque réplica présente de légères particularités qui témoignent d'un comportement dynamique différent bien que la structure initiale et les conditions de simulation soient identiques. Ces différences ne consistent pas toujours en l'apparition ou disparition de nouvelles structures secondaires, mais davantage en une variation de la fréquence d'apparition de ces structures – leur occurrence. Ainsi, les deux réplicas de STAT5a (*cf.* Figure 37) ne diffèrent significativement que par la fréquence d'occurrence du ponts β dans le DBD (résidus 427 – 430) ainsi que par l'extension d'un pont β entre les résidus 669 et 673 en un feuillet β entre les résidus 668 - 669 et 673 - 674. D'autres changements apparaissent entre les réplicas de STAT5a phosphorylé : la présence d'un pont β entre les résidus 356 – 359, l'apparition d'un pont/feuillet β (résidus 394 – 395), un équilibre hélice α/3_{10} différent pour les hélices αA (résidus 599 - 605) et αD (résidus 674 - 681). Les différences entre les deux réplicas de STAT5b consistent en l'apparition d'hélices α et 3_{10} dans le domaine de liaison à l'ADN (résidus 387 – 389 et 395-397), une hélice *α5* moins présente (résidus 472 – 475) et la quasi disparition de l'hélice *αD* dans le second réplica. Enfin les réplicas de pSTAT5b montrent un changement de structures des résidus 356 – 359, formant un pont β dans le premier réplica, et impliqué dans une hélice 3_{10} dans le second réplica, une hélice α dans le DBD présent uniquement dans le second réplica (résidus 375-379), un pont β (résidus 382 – 385) dans le premier réplica, des structures complètement différentes au niveau de la queue phosphotyrosyl (présence d'une hélice 310 au niveau des résidus 690 – 693 et d'un pont β entre les résidus 702 et 705).

Différences observées entre les deux isoformes de STAT5, STAT5a et STAT5b :

Plusieurs différences entre les deux isoformes de STAT5 peuvent être notées en termes de structures secondaires. L'hélice *α2* est ainsi systématiquement plus longue à son extrémité N-terminale dans les simulations de STAT5b ou pSTAT5b tandis que l'extrémité de l'hélice α1 est davantage structurée en hélice 3_{10} dans les simulations de (p)STAT5b (*cf.* Figure 37). La transformation d'une forme d'hélice en une autre (α → 3_{10}) entre les deux isoformes correspond parfaitement aux variations de deux résidus généralement favorables à une structuration en

hélice (A187 et Q188 chez STAT5a) vers des résidus peu favorables aux hélices (G187 et P188). Les effets observés dans cette région peuvent ainsi être expliqués par ce polymorphisme de la séquence peptidique. Les brins *b* et *e* sont légèrement raccourcis dans STAT5b et pSTAT5b, au niveau du résidu modifié V442I qui se situe dans le brin *f*, positionné à proximité. L'encombrement légèrement supérieur de la chaîne latérale du résidu isoleucine vient en effet déstabiliser le feuillet β formé par les brins *b* et *e*.

La présence de nombreuses variations de résidus dans le domaine SH2 se traduit par la disparition du brin *C* du feuillet β dans les isoformes phosphorylés et non-phosphorylés de STAT5b. Quatre résidus modifiés sont situés à proximité du brin *C*: 3 sont situés dans la boucle reliant les brins *B* et *C* (P636Q, N639M et L640F) tandis que la variation du résidu K644M se trouve au milieu du brin *C*. Le résidu F679 (STAT5a) semble également favoriser le maintien de l'hélice *αD*, à laquelle il appartient, comparativement à son homologue Y679 (STAT5b), qui entraine une déstabilisation de cette hélice. Enfin, la présence de l'insertion de cinq résidus (CESAT) dans l'isoforme b de STAT5 favorise la stabilisation de plusieurs structures secondaires supplémentaires, alors qu'aucune structure secondaire n'est observée dans l'isoforme a au niveau de la queue (phospho-)tyrosyl. La présence de résidus modifiés dans deux séquences et de changements structuraux associés nous amène à proposer que les différence de séquence observées entre STAT5a et STAT5b engendrent les effets locaux décrits ci-dessus, à savoir la modification des extrémité C-terminale de l'hélice *α1* et de l'extrémité de l'hélice *α2*, la modification de structures au sein du DBD se traduisant par le raccourcissement des brins *b* et *e*, la disparition du brin *C* dans le domaine SH2, la déstabilisation de l'hélice *αD* et l'établissement de structures secondaires propres dans la queue (phospho-)tyrosyl.

Malgré leur grand taux d'identité de séquence, l'analyse détaillée des caractéristiques des structures secondaires des deux isoformes, STAT5a et STAT5b, a révélé des différences notables entre chaque paire de simulations de DM. Ces changements dénotent l'influence des perturbations induites par les changements d'acides aminés liés à la divergence des deux isoformes, et reflètent les réarrangements structuraux et des dynamiques conformationnelles dépendantes de la séquence. Les changements les plus notables sont observés à proximité directe des sites de polymorphisme (brin *C* dans le domaine SH2, brins *b* et *e* du DBD, hélice *α1*, …).

Le domaine SH2 est fonctionnellement primordial pour les protéines STAT5 puisqu'il est impliqué dans la reconnaissance de protéines partenaires, un rôle critique pour les protéines impliquées dans la signalisation cellulaire. Ce domaine sert ainsi d'interface avec les protéines contenant des résidus de phosphotyrosine, à commencer par STAT5 lors de la formation des dimères parallèles. De manière plus générale, le domaine SH2 associe différentes protéines des voies de signalisation contenant une phosphotyrosine afin de recruter spécifiquement une protéine, de former des assemblages multi-protéiques ou de participer à la régulation d'activités protéiques[596,597]. Il a été montré que la régulation des protéines phosphorylées peut se faire à la fois par la régulation de la phosphorylation (mutation du site de phosphorylation) mais également par la modification de l'environnement du site de phosphorylation [600]. Dans ce cas,

la modification du site de reconnaissance du résidu de phosphotyrosine (le domaine SH2) pourrait constituer une expression de la différenciation des isoformes de STAT5, *via* la modulation de l'affinité entre les différentes protéines STATs lors de la formation d'hétérodimères impliquant STAT5.

L'impact des modifications structurales liées aux changements de séquence dans le DBD peut être abordé en termes de changement de l'affinité des sites de liaison à l'ADN. La perturbation des structures et de la dynamique de ce domaine module potentiellement la reconnaissance de la protéine pour le site ADN ciblé, par exemple un site promoteur sur lequel STAT5 se fixe. Cependant, il n'existe pas à ce jour de preuves expérimentales mettant en évidence une différence d'affinité pour l'ADN entre STAT5a et STAT5b [524]. L'impact fonctionnel des changements structuraux et leurs rôles dans la régulation de l'activité de STAT5 restent donc à établir.

Le domaine CCD est impliqué dans la régulation de l'import nucléaire des protéines STATs, et plus particulièrement pour STAT3, STAT5 et STAT6[115]. Précisément, STAT5 nécessite la présence des 4 hélices de la partie proximale du CCD pour être capable de passer du compartiment cytoplasmique au compartiment nucléaire. De plus, Shin et Reich ont récemment cartographié les résidus nécessaires à l'import de STAT5 dans le noyau, et ont montré que les résidus essentiels au trafic 'cytoplasme – noyau' sont les résidus E149 (hélice *α1*), R241, K242, R257, R258 (hélice *α2*) et I320 (hélice *α4*)[113]. Ainsi, l'arrangement local du CCD proximal est nécessaire à la fonction de STAT5. Le CCD distal ne semble pas impliqué dans cette fonction d'import, ni dans d'autres fonctions. Il est néanmoins impossible d'établir un lien entre les deux résidus modifiés entre les deux isoformes de STAT5 (A187G et Q188P), les changements de structures (changement de la structuration en hélice α vers une hélice 3_{10}) et un rôle fonctionnel. De même, la capacité d'oscillation du CCD distal par rapport au CCD proximal n'a jamais été mentionnée dans la littérature jusqu'à présent, et constitue un élément nouveau qui nécessite davantage d'études afin de comprendre son rôle éventuel dans la régulation de STAT5. Dans les autres protéines STATs, ces éléments ne sont pas présents du fait d'une longueur des hélices α1 er α2 plus courtes.

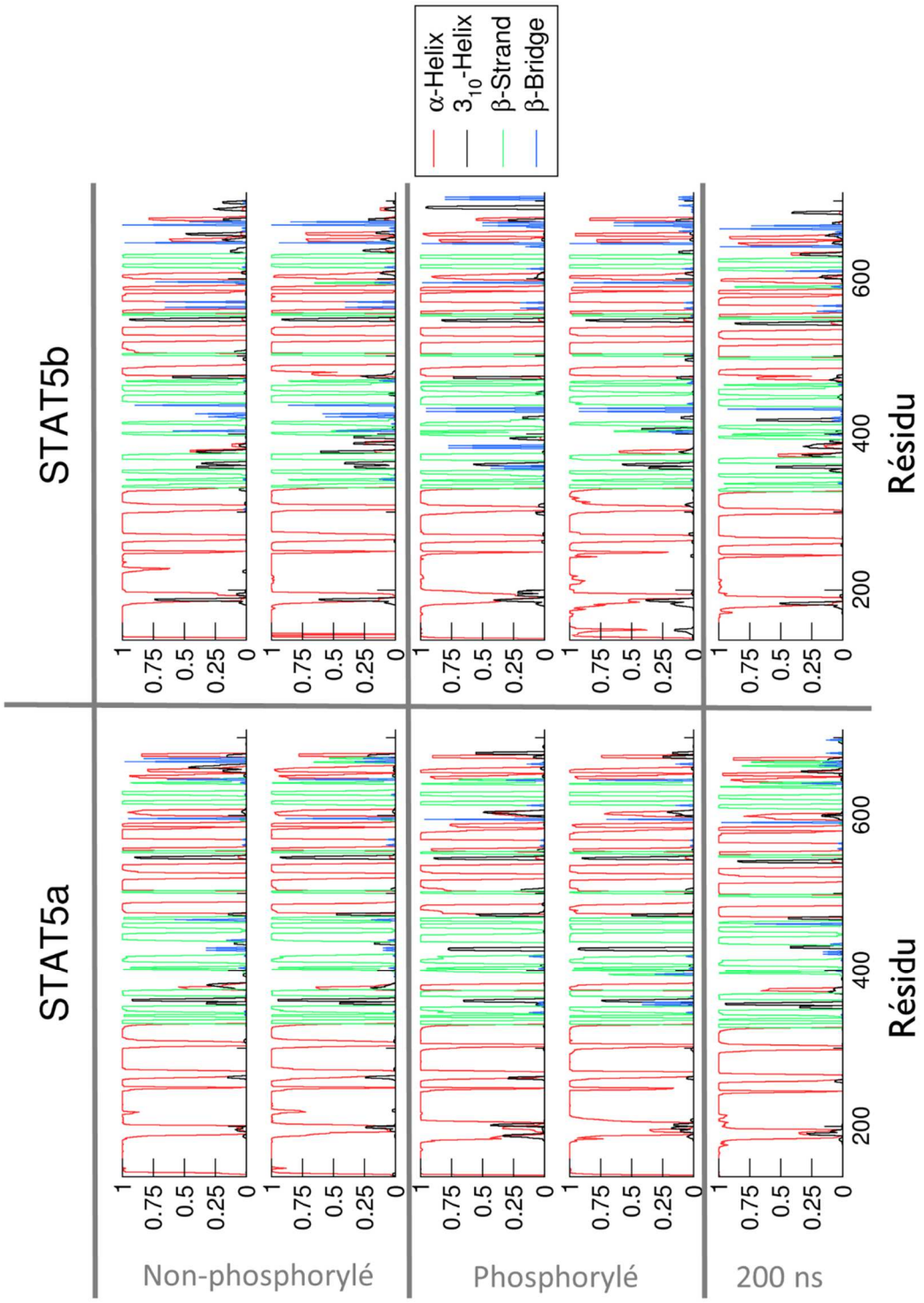


Figure 37: **Variations des structures secondaires au cours des simulations de dynamique moléculaire.** Les simulations de STAT5a sont à gauche, et de STAT5b à droite. Les simulations se rapportant aux systèmes non-phosphorylés sont dans les cadrans supérieurs, aux systèmes phosphorylés dans les cadrans centraux et aux dynamiques étendues dans les cadrans inférieurs. Les structures secondaires sont exprimées en probabilité d'existence au cours de la simulation.

Différences observées entre les formes phosphorylées et non-phosphorylées de STAT5:

De manière similaire, des différences peuvent être observées entre les formes phosphorylées et non phosphorylées des deux isoformes de STAT5 (*cf.* Figure 37). Chez STAT5a, la présence du groupement phosphate se traduit par la disparition de brins/ponts β dans le domaine SH2, ces brins étant bien présents dans la structure initiale à l'issue de la modélisation par homologie. On peut donc avancer que la présence du groupement phosphate déstabilise ces structures. Le même effet est retrouvé, de manière moins marquée cependant, chez STAT5b, où des ponts β sont moins présents dans les simulations de la protéine phosphorylée (résidus 668 et 672). Un autre effet local est observé chez STAT5b, à savoir la plus faible stabilisation des hélices *αB* et *αC* illustrée par leur occurrence plus modérée. À ces effets locaux, d'autres effets à plus longue portée peuvent être notés. La présence quasi-constante d'une hélice 3_{10} au sein du DBD (résidus 427 à 430) dans les formes phosphorylées de STAT5a comparativement aux formes non-phosphorylées peut être une manifestation des effets longue portée de la liaison du groupement phosphate sur le résidu 694. Nous observons également un effet de phosphorylation chez STAT5b, mais qui se manifeste différemment. Un pont β est présent de manière transitoire entre les résidus 422 et 426 dans les simulations de STAT5b non phosphorylé, tandis qu'aucune structuration de cette région n'est observée dans les formes phosphorylées. La présence d'un effet longue portée du groupement phosphate qui impacte la même région de manière différente est importante. Il s'agit de la manifestation d'un couplage que l'on peut qualifier d'allostérique. Cependant, la fonction d'un tel couplage reste peu claire, étant donné que les résidus du DBD évoqués n'ont jamais été relevés comme étant cruciaux dans la régulation et la fonction de STAT5.

Les évènements de phosphorylation/déphosphorylation sont courants dans le milieu cellulaire et participent de manière continue à la régulation des processus physiologiques ou physiopathologiques. En général, la liaison d'un groupement phosphate va permettre un gain de fonction de la protéine *via* l'apport d'un groupement hautement polaire et chargé. La phosphorylation modifie ainsi les voies de signalisation activées en modifiant le paysage des interactions protéine – protéine, mais peut avoir également d'autres effets, comme des changements conformationnels[601,602], de la cinétique[603], ou la régulation de la localisation cellulaire[604,605]. Les protéines STATs font partie des protéines dont la fonction est affectée par ce type d'évènements, soit au niveau de la tyrosine 694/699 (chez STAT5a/b), soit au niveau des sérines du domaine TAD ou encore par la présence du domaine SH2. Nous avons vu que la phosphorylation d'un résidu ne modifie pas l'architecture général de STAT5, mais induit des modifications ou arrangements structuraux à courte portée au niveau de la queue phosphotyrosyl qui stabilise la formation de structures secondaires dans son environnement, comme observé dans le cas d'une autre protéine phosphorylée, le récepteur à l'acide rétinoïque α[606]. En plus de ces effets locaux, des effets à longue portée ont été notés au niveau du domaine de liaison à l'ADN, qui se manifestent sur la même région (résidus 422 – 430), mais de manière différente dans les deux formes de la protéine. La forme phosphorylée de STAT5a montre la stabilisation d'une hélice, alors que chez STAT5b phosphorylé, on peut voir la disparition

complète d'un pont β transitoire et l'établissement d'un autre pont β stable, résultant en une rigidification globale de cette région. L'absence de données structurales pour les formes phosphorylées des protéines STATs ne permet pas de dire si cette propriété est partagée par l'ensemble de la famille des STATs.

B. Analyse des mouvements collectifs des STAT5

Afin de compléter la description du paysage dynamique des protéines STAT5, nous avons réalisé l'analyse des mouvements globaux. Pour cela, nous avons calculé les modes principaux (*cf.* paragraphe II.F.1 du chapitre 2) à partir de la matrice de covariance entre les atomes de carbone α des protéines. Il a été montré que les composantes principales issues de simulations dont les vitesses initiales sont différentes produisent des vecteurs qui divergent de manière significative (caractérisé par un recouvrement inférieur à 0,7) [607,608]. L'espace conformationnel du système n'est donc échantillonné que de manière partielle par chaque simulation. Nous avons fait la même observation pour STAT5, quel que soit l'isoforme et la phosphorylation (*cf.* Tableau 7).

Tableau 7: **Recouvrement des cinq premiers modes de l'ACP de chaque paire de dynamique.** Les valeurs correspondant à STAT5a sont en bleu, à pSTAT5a en jaune, à STAT5b en vert et à pSTAT5b en magenta. Une valeur de 1 correspond à un recouvrement total des modes, une valeur de 0 indique l'orthogonalité des vecteurs. Les recouvrements sont indiqués en valeur absolue.

	Mode 1	Mode 2	Mode 3	Mode 4	Mode 5
Mode 1	0.65 0.17 0.17 0.36	0.25 0.19 0.18 0.38	0.21 0.37 0.34 0.25	0.21 0.28 0.23 0.24	0.10 0.06 0.04 0.13
Mode 2	0.04 0.08 0.49 0.01	0.25 0.20 0.25 0.00	0.38 0.56 0.39 0.54	0.10 0.54 0.13 0.00	0.23 0.13 0.21 0.13
Mode 3	0.14 0.56 0.16 0.17	0.55 0.09 0.65 0.25	0.41 0.14 0.03 0.31	0.27 0.42 0.16 0.05	0.20 0.10 0.39 0.10
Mode 4	0.19 0.22 0.15 0.30	0.26 0.06 0.08 0.03	0.05 0.29 0.16 0.25	0.14 0.07 0.46 0.26	0.44 0.55 0.06 0.13
Mode 5	0.50 0.14 0.32 0.21	0.17 0.29 0.40 0.14	0.04 0.02 0.04 0.04	0.35 0.14 0.07 0.33	0.14 0.27 0.39 0.11

Des points communs sont observés entre tous les modes principaux des différentes simulations de dynamique moléculaire des différentes STAT5. Tout d'abord, les régions de STAT5 qui montrent les plus amples mouvements sont d'une part le CCD distal et d'autre part la queue (phospho)-tyrosyl, ce qui corrèle bien avec les profils de RMSFs (*cf.* Figure 36 et Figure 38). Le premier mode révèle cependant des différences de comportement entre les simulations de DM, à la fois dans l'amplitude des mouvements des deux segments mais également dans la direction des fluctuations principales. Ces observations sont cohérentes avec les faibles valeurs de superposition des modes principaux évoquées dans le paragraphe précédent et décrites dans la littérature pour d'autres systèmes[607,608].

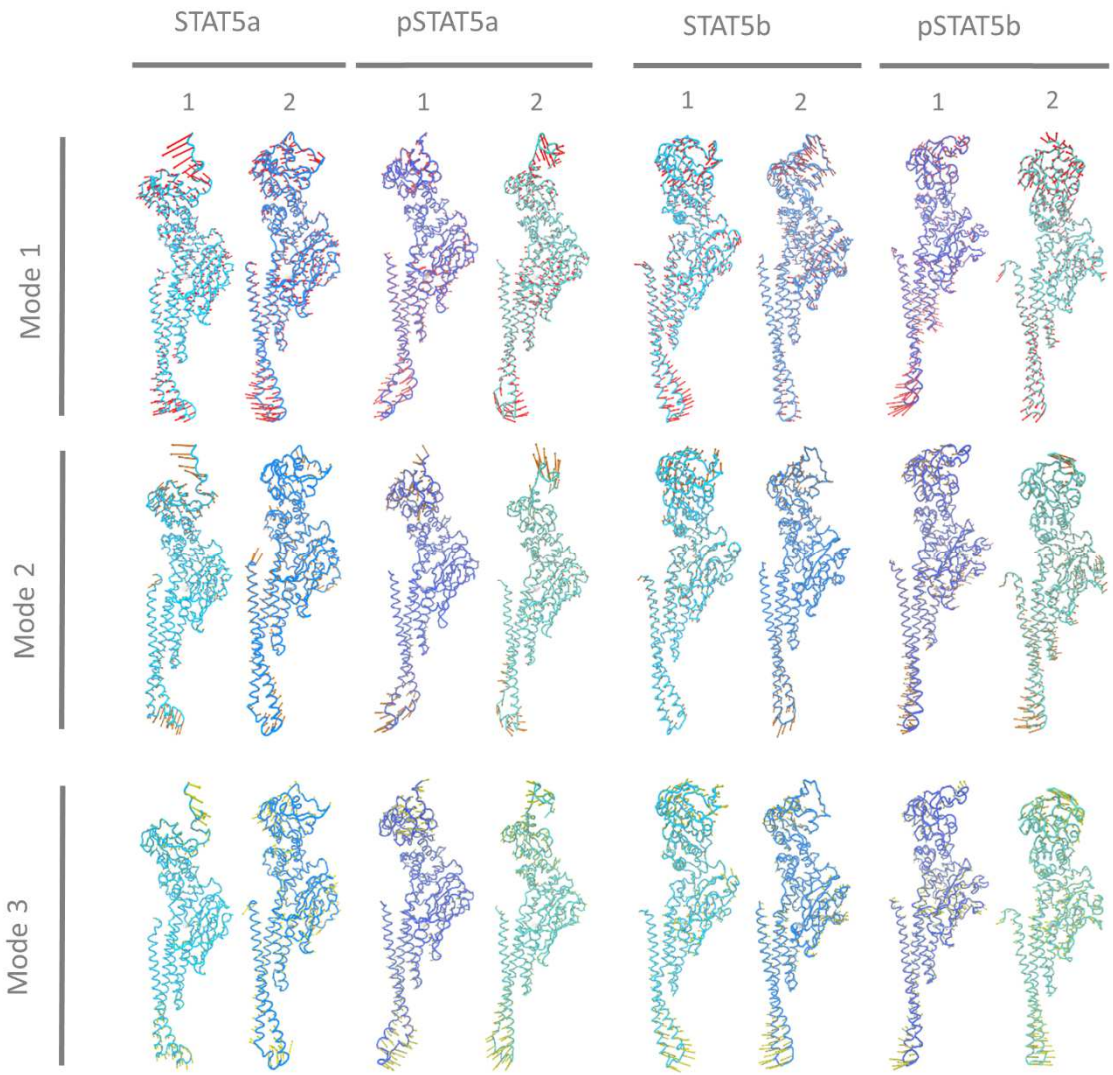


Figure 38: **Mouvements globaux de STAT5 calculés par analyse des composantes principales des simulations de dynamiques moléculaires.** Les modes de STAT5a, pSTAT5a, STAT5b et pSTAT5b sont montrés de gauche à droite. Les premier, second et troisième modes sont présentés en haut, au milieu et en bas, respectivement.

Ces mouvements des extrémités distales de STAT5 s'accompagnent d'autres mouvements de moindre amplitude au niveau du domaine de liaison à l'ADN et du domaine

SH2. Ce dernier montre un déplacement similaire, quoique de plus faible amplitude, à celui de la queue (phospho-)tyrosyl, à l'exception de la seconde réplique de DM de pSTAT5a et de la première réplique de pSTAT5b. Le domaine DBD semble adopter un mouvement de rotation autour de l'axe qui passe à travers le CCD proximal et le domaine LD. Ces deux domaines présentent les plus faibles mouvements en termes d'amplitude (*cf.* Figure 39).

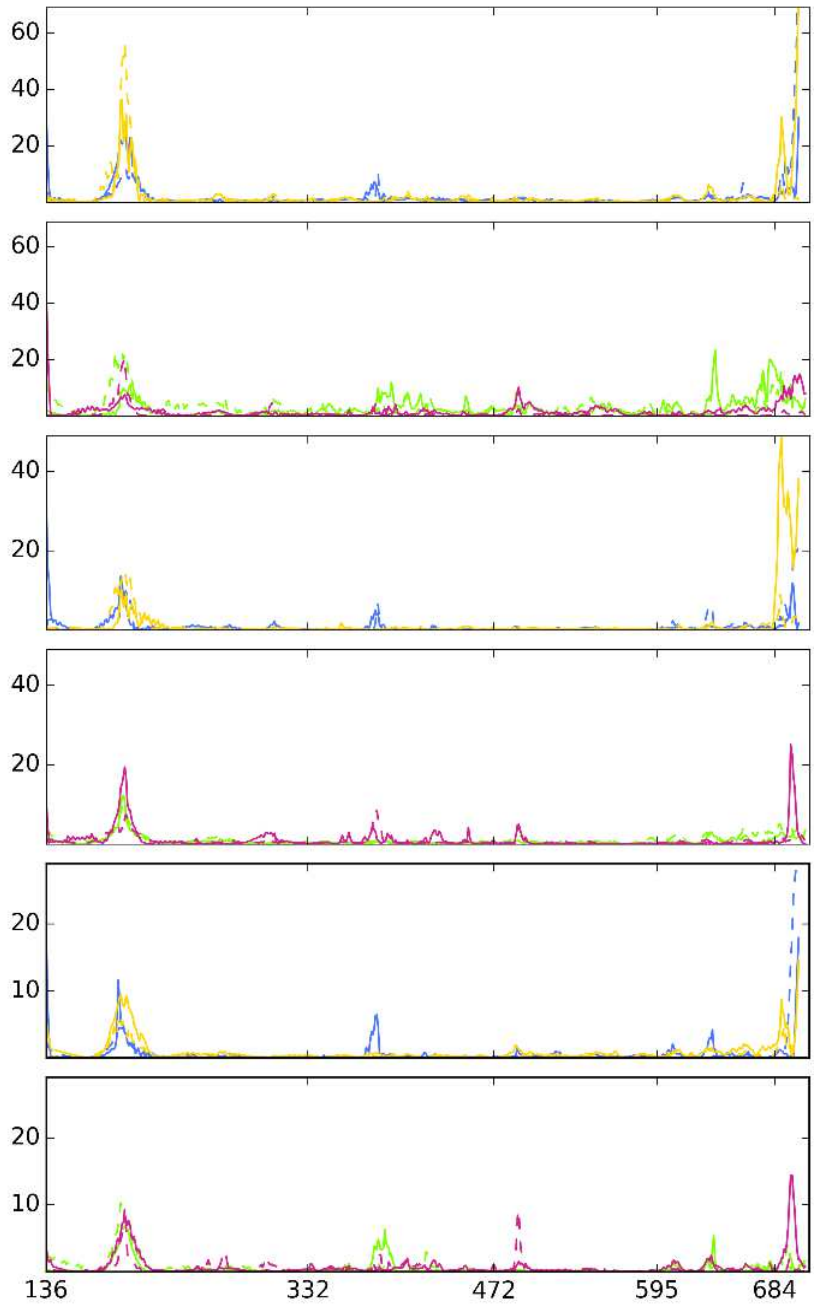


Figure 39: **Analyse en composante principale des simulations de dynamique moléculaire de STAT5. Fluctuations des 3 premiers modes ACP de chaque simulation.** Les simulations de DM de STAT5a sont en bleu (trait plein et pointillé), de pSTAT5a en jaune (trait plein et pointillé), de STAT5b en vert (trait plein et pointillé) et de pSTAT5b en magenta (trait plein et pointillé).

Le deuxième et le troisième mode présentent également des mouvements corrélés mais de manière moins marquée que le premier mode. Selon ces modes, les régions de STAT5 les plus

mobiles restent la queue (phospho-)tyrosyl et le CCD distal, alors que le domaine SH2 et le DBD sont très peu mobiles. Le DBD présente néanmoins de fluctuations plus importantes dans le second mode de pSTAT5b, comparativement aux autres systèmes, alors que dans le troisième mode, STAT5b et pSTAT5b montrent des variations du DBD plus importantes que STAT5a et pSTAT5a. Enfin, les mouvements du CCD distal et de la queue (phospho-)tyrosyl sont décrits dans tous les systèmes par les trois premiers modes.

Le premier mode explique à lui seul entre 18,8 (pSTAT5b) et 48,2% (pSTAT5b) de la variance totale du système analysé (*cf.* Figure 40). Les deux dynamiques de STAT5b sont celles qui présentent le premier mode le plus important (celui expliquant la plus grande fraction de variance). Cependant, ces deux modes ont un recouvrement entre eux très faible (0,17, *cf.* Tableau 7), mais également avec les autres modes. Les transitions décrites par les premiers modes sont donc propres à chaque système.

La comparaison des modes principaux dans chaque système ne révèle pas de différences significatives entre STAT5a et STAT5b, phosphorylés ou non phosphorylés. Ainsi, les deux isoformes ont une dynamique globale quasiment identique, que ce soit au niveau des régions qui présentent des mouvements concertés ou des amplitudes de ces mouvements. STAT5b et pSTAT5b montrent cependant de plus larges variations du DBD dans les modes 2 (pSTAT5b) et 3 (STAT5b et pSTAT5b). Aucune différence notable n'a pu être extraite concernant l'impact de la phosphorylation sur la dynamique globale des protéines.

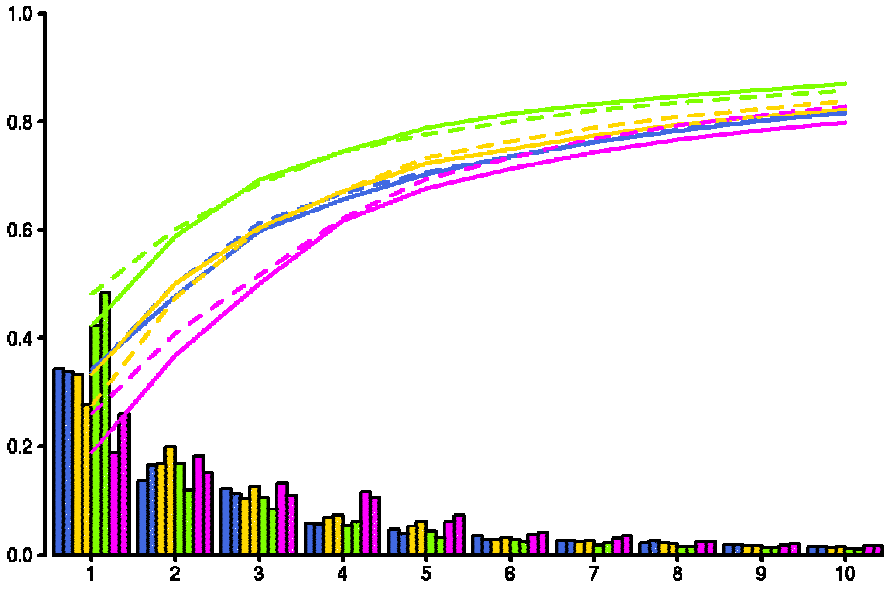


Figure 40: **Fraction de variance du système expliquée par chaque mode d'ACP.** Les histogrammes indiquent la contribution de chaque mode, les courbes indiquent la contribution cumulée des modes. STAT5a est en bleu, pSTAT5a est en jaune, STAT5b est en vert et pSTAT5b est en magenta. Les deux réplicas de chaque protéine sont différenciés par le type de trait de la courbe (plein ou pointillé) ou le remplissage des barres d'histogramme (plein ou hachuré).

C. Mouvements harmoniques de STAT5

Nous avons calculé les modes normaux à partir des structures équilibrées (les structures à t = 0 ns) afin de déterminer si des mouvements de basse fréquence apparaissent par rapport aux mouvements observés dans l'analyse par composante principale. Les différences entre les isoformes seront également étudiées afin de déterminer l'influence du polymorphisme de séquence sur la dynamique intrinsèque de STAT5.

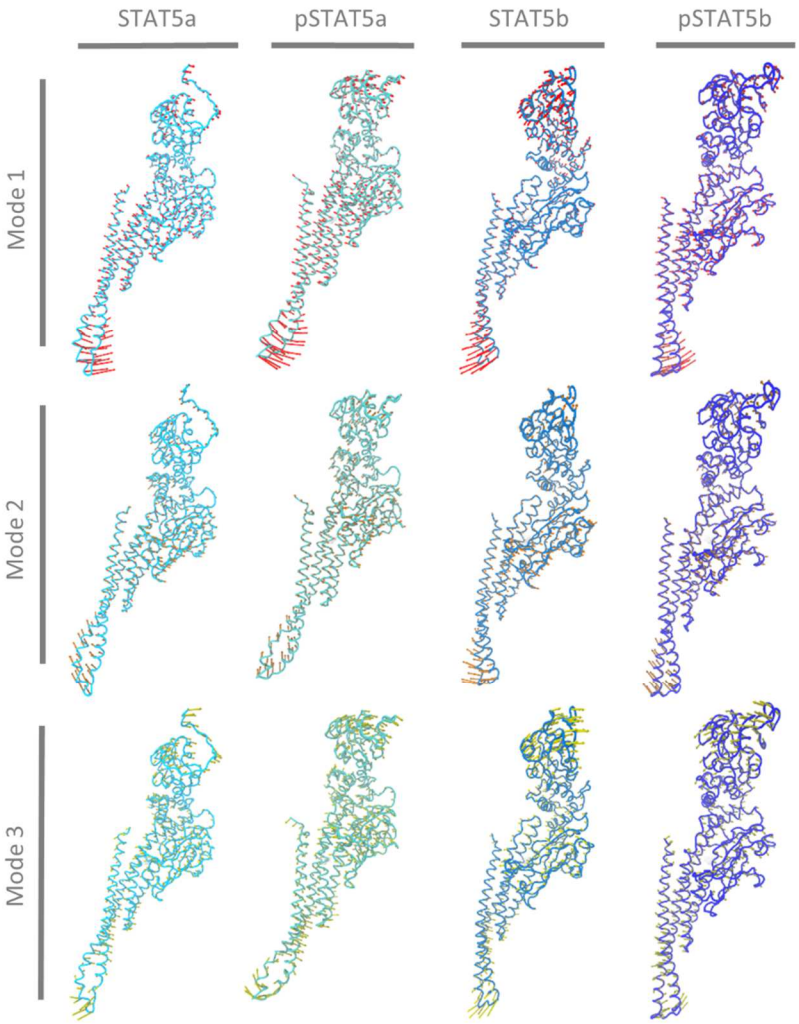


Figure 41: **Modes normaux des monomères de STAT5.** Pour chaque espèce, les trois premiers modes normaux sont affichés.

Le CCD distal est la région de la protéine qui présente les mouvements les plus amples (*cf.* Figure 41), dans des directions similaires pour toutes les protéines et pour le mode de plus basse fréquence (Mode 1). L'amplitude des mouvements du CCD distal diminue rapidement, et l'amplitude des mouvements du domaine SH2 et de la queue (phospho-)tyrosyl deviennent du même ordre. La queue (phospho-)tyrosyl, au contraire des modes issus de l'analyse en composante principale, montre des mouvements similaires à ceux observés dans le domaine SH2, en termes de direction et d'amplitude. Cette différence de comportement dynamique

s'explique par la méthodologie différente des deux approches, analyse en composante principale et analyse des modes normaux. La seconde ne prend pas en compte les effets de solvatation ni les mouvements non-harmoniques. Les fluctuations du domaine SH2 montrent une plus grande flexibilité des boucles liant les différents éléments structuraux. Entre deux régions éloignées (CCD distal et domaine SH2 – queue (phospho-)tyrosyl), le domaine DBD et le CCD proximal montrent une dynamique commune, caractérisée par des mouvements similaires et de même amplitude. Le LD est partagé entre les influences des domaines environnants, la partie proche du DBD et du CCD proximal partage le même type de mouvements, alors que l'extrémité proche du domaine SH2 adopte un comportement similaire. Contrairement aux boucles du domaine SH2, les boucles du domaine de liaison à l'ADN ne présentent pas de fluctuations différentes du reste du DBD dans le premier mode ; une premier pic autour du résidu 385 (boucle reliant les brins c et c', *cf.* Figure 42 milieu) est cependant observé pour tous les systèmes pour le second mode normal, mais n'est plus observé pour les autres modes.

À l'inverse des modes issus de l'ACP, les modes normaux se recouvrent très bien, entre protéines étudiées. Les matrices des produits scalaires montrent une excellente correspondance (caractérisée par des valeurs de recouvrement supérieures à 0,7) entre les modes normaux des protéines phosphorylées et non-phosphorylées. Chez STAT5a, l'ordre des modes est conservé entre les modes 1– 4, 7 et 10, alors que le mode 6 de la forme non-phosphorylée correspond au mode 5 de la forme phosphorylée. Les modes 8 à 10 montrent uniquement des recouvrements faibles (*i.e.* 0,5). Les modes des deux formes de STAT5b produisent une diagonale plus nette, les modes 1 à 3 et 6 à 8 se superposant très bien. Les autres modes (4 – 5 et 9 – 10) présentent un recouvrement partiel (0,4 à 0,6) à l'exception du mode 10 de la forme non-phosphorylée qui correspond au mode 9 de la forme phosphorylée.

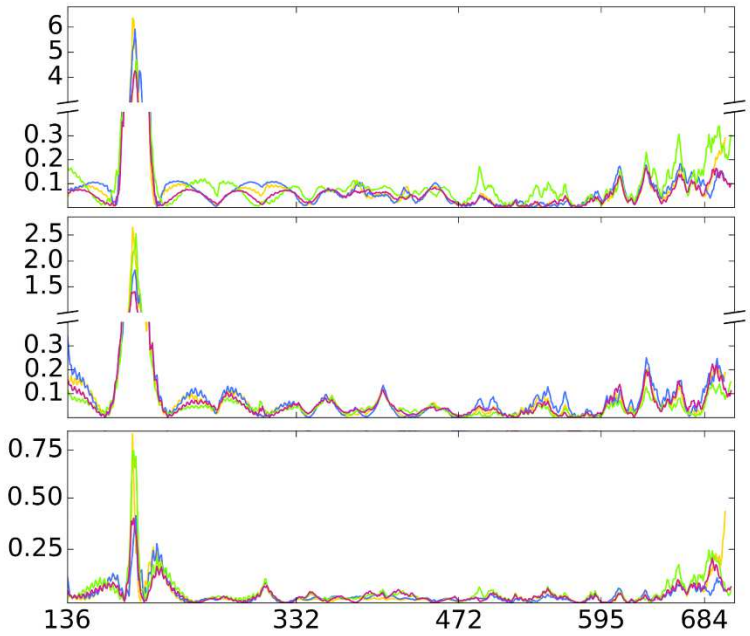


Figure 42: **Mouvements globaux de STAT5.** Fluctuations des carbones α pour les trois premiers modes normaux. Les fluctuations sont redimensionnées pour avoir la même moyenne ; l'échelle des fluctuations est donc arbitraire. STAT5a est en bleu, pSTAT5a en jaune, STAT5b en vert et pSTAT5b en magenta. Le mode 1 est dans le panneau du haut, le mode 2 dans le panneau intermédiaire et le mode 3 dans le panneau du bas. L'axe des abscisses représente l'indice des résidus.

Les mouvements observés par l'analyse des modes normaux sont très amples au niveau du CCD distal à l'instar de ceux observés par l'analyse en composante principale, mais plus reproductibles d'une protéine à l'autre. Ils s'accompagnent d'un autre mouvement relativement ample au niveau du domaine SH2, auquel est attaché la queue (phospho-)tyrosyl, qui présente des mouvements beaucoup moins amples que ceux observés lors de l'ACP. Enfin, le DBD et le CCD proximal se comportent comme un ensemble dynamique homogène. Comme l'ACP l'a montré, les dynamiques globales des deux isoformes de STAT5 et des formes non-phosphorylées et phosphorylées sont comparables. Le rôle des mouvements du CCD distal reste une question ouverte, car il n'a jamais été évoqué à notre connaissance.

D. Corrélation des mouvements de STAT5

Une représentation générale des mouvements collectifs peut être obtenue en calculant les corrélations croisées des fluctuations atomiques. Les résidus qui se déplacent dans la même direction sont corrélés, alors que ceux qui se déplacent dans la direction opposée sont anti-corrélés ou corrélés négativement. Afin d'accroitre notre connaissance des mouvements de STAT5, nous avons comparé les motifs issus des cartes de corrélation croisée générés par les modes normaux et par les modes de l'ACP. Nous avons appliqué cette analyse afin d'identifier les effets de couplage dynamique longue portée, qui sont associés aux phénomènes de régulation fonctionnel des protéines ou de dérégulation provoquant le dysfonctionnement des protéines[609,610].

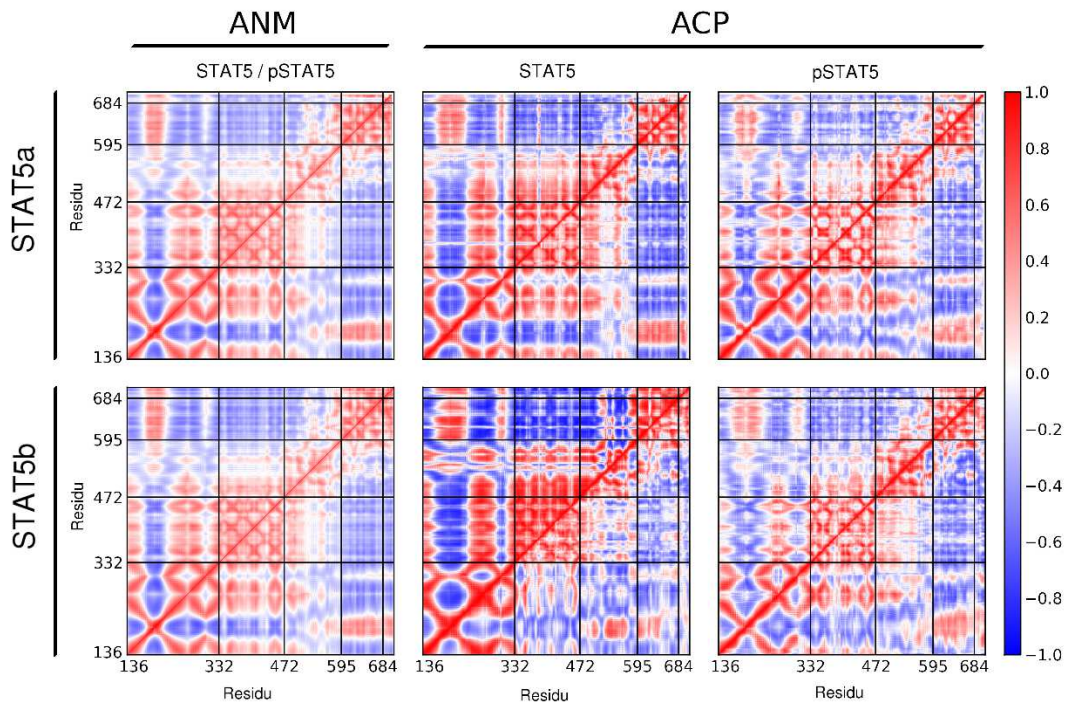


Figure 43: Corrélations croisées de STAT5. Les cartes issues des modes normaux sont à gauche, STAT5a/pSTAT5a en haut et STAT5b/pSTAT5b en bas. Chaque réplica de dynamique (30 ns) analysé par ACP (colonnes du milieu et de droite) est présenté dans la moitié supérieure ou inférieure. Les corrélations sont représentées par un gradient de couleur variant du bleu (négative) ou rouge (positive) en passant par le blanc (absence de corrélations).

Les cartes des corrélations croisées ont été calculées pour tous les modes, issus soit de l'analyse des modes normaux (colonne de gauche), soit de l'analyse en composante principale (colonnes centrale et de droite) (*cf.* Figure 43). Les cartes des modes normaux indiquent que dans les deux protéines, STAT5 (moitié supérieure des cartes) et pSTAT5 (moitié inférieure des cartes), les motifs de corrélations sont similaires et indiquent des mouvements couplés entre des sites distants, en particulier la région distale du CCD (correspondant aux résidus 180 à 220) et le domaine SH2 (résidus 595 à 684), pourtant séparés par plus de 80 Å. Le CCD distal montre des mouvements anti-corrélés avec le CCD proximal (constitué de la partie N-terminale de l'hélice α1, de la partie C-terminale de l'hélice α2 et des hélices α3 et α4), le DBD et la partie du LD la plus proche du DBD, indiquant des mouvements corrélés dans le sens opposé. De plus, le CCD distal présente des mouvements hautement corrélés avec le domaine SH2 et la queue (phospho-)tyrosyl, qui montrent des mouvements anti-corrélés avec le CCD proximal et le DBD. De telles caractéristiques de corrélations peuvent être expliquées par l'architecture commune des protéines STATs, qui ont une forme générale allongée ou tubulaire. Les mouvements d'une extrémité (le CCD distal) sont contrebalancés par les mouvements de l'extrémité opposée (le domaine SH2 et la queue (phospho-)tyrosyl) afin que la protéine soit dans une forme d'équilibre stable autour de son centre de gravité. Une telle régulation pourrait être soumise à des phénomènes allostériques.

Les cartes de corrélations croisées calculées à partir des ACPs de chaque simulation de dynamique moléculaire montrent des caractéristiques similaires pour toutes les protéines STAT5 étudiées, et correspondent aux cartes de corrélations croisées issues de l'analyse des modes normaux. Les principales caractéristiques des corrélations dynamiques des protéines STAT5 non-phosphorylées sont d'une part la taille des fragments hautement corrélés (positivement ou négativement), plus longs dans les domaines CCD, DBD et LD, et plus courts dans le domaine SH2 et d'autre part l'augmentation des valeurs des corrélations comparativement à cartes des modes normaux. De manière similaire aux protéines non-phosphorylées, l'analyse des corrélations de pSTAT5a et pSTAT5b montrent des mouvements fortement corrélés entre des sites distants. Néanmoins, la phosphorylation induit une légère diminution des corrélations (positives et négatives) dans les deux isoformes de STAT5, indiquant une diminution du couplage des mouvements entre domaines (CCD et SH2) qui pourrait affecter la spécificité des sites de liaison de la protéine pour les partenaires cellulaires.

L'analyse détaillée des corrélations dynamiques de pSTAT5a et pSTAT5b, ainsi que de celles de STAT5a et STAT5b, a montré une très bonne similarité du comportement des protéines. Dans les détails, sous l'effet de la phosphorylation, nous avons observé que les corrélations et anti-corrélations de tous les domaines sont légèrement diminuées, et se manifestent par un degré de corrélation globalement plus faible, ainsi que par une variation de la taille des fragments positivement ou négativement corrélés.

E. Dynamique locale de STAT5 et chemins de communication étudiés par MOdular Network Analysis - MONETA

La caractérisation des conformations de STAT5 issues des simulations de dynamique moléculaire et de leurs structures secondaires a mis en avant des variations spécifiques à chaque protéine. Notre hypothèse est que ces effets structuraux sont reliés à la fois à la séquence des protéines et à leur statut de phosphorylation. Les changements structuraux induits par les différences d'acide aminé montrent des modifications localisées à proximité de ces sites de polymorphisme. L'impact de la phosphorylation de STAT5 a révélé des effets locaux, mais également longue-distance. Les corrélations croisées issues de l'ACP et de l'AMN ont mis en évidence des mouvements fortement corrélés entre des domaines distants de la protéine. Afin de comprendre l'origine de ces changements structuraux à partir des variations de séquences et/ou présence/absence du groupement phosphate, nous avons caractérisé les propriétés dynamiques locales de toutes les protéines STAT5s et analysé les chemins de communication intra-protéique, pour repérer des réseaux d'interactions reliant des sites distants. Afin d'examiner ces caractéristiques, nous avons utilisé l'Analyse en Réseau Modulaire, MONETA (*MOdular NETwork Analysis*), une méthode appliqué précédemment à l'étude de la communication allostérique dans les récepteurs à activité tyrosine kinase[568–570], ainsi qu'une nouvelle approche que nous avons développé en collaboration avec le CMLA, la Décomposition en Traits Principaux (*Principal Feature Decomposition*), conçue pour caractériser la dynamique locale des résidus.

1. Identification des Segments Dynamiques Indépendants (IDSs)

Comme première étape de la caractérisation des protéines par MONETA, les régions de STAT5 présentant les propriétés dynamiques internes les plus marquantes ont été identifiées par une technique statistique appelée « Analyse en traits locaux » (*LFA, Local Feature Analysis*)[572] adapté à l'étude de la dynamique essentielle des protéines[573,574]. Ce formalisme permet d'identifier des résidus « graines » dans les systèmes étudiés et de définir ensuite des groupes de résidus spatialement proches et ayant des fluctuations atomiques proches. Des groupes de résidus partageant une dynamique commune à l'échelle locale sont ainsi définis, et dénommés « Segments Dynamiques Indépendants » (*IDSs, Independent Dynamics Segments*) car ils présentent une dynamique interne quasi-indépendante de celle du reste de la protéine.

Dans les protéines STAT5a phosphorylées et non-phosphorylées, le nombre d'*IDSs* défini à partir des résidus « graines » est identique et la position des 14 *IDSs* se superposent très bien (*cf.* Figure 44 C). Nous avons attribué un numéro *i* allant de 1 à 14 à chaque segment *S*, en fonction de leur ordre d'apparition dans la séquence de STAT5a. Les *IDSs* non-présents dans STAT5a mais présents dans STAT5b ont été numérotés de 15 à 24 par la suite afin de permettre leur distinction. Dix *IDSs* (S1, 3-11) de STAT5a ont un caractère discret, les quatre autres (S2, 12-14) montrent un chevauchement partiel, et peuvent être interprétés comme des *IDSs* fusionnés ou dupliqués. Les *IDSs* de STAT5b montrent de larges différences à la fois en nombre et en caractère : les 9 *IDSs* identifiés dans STAT5b sont distincts et séparés, alors que les 24 *IDSs* de pSTAT5b se superposent partiellement et peuvent être vus comme 13 *IDSs* fusionnés.

Dans STAT5a, les 14 *IDSs* identifiés sont distribués sur l'ensemble des domaines de manière homogène (*cf.* Figure 44 A, C). Ils sont principalement retrouvés dans les régions flexibles, et peuvent intégrer des fragments plus rigides proches dans l'espace. Dans le CCD, trois *IDSs* sont localisés à l'extrémité N-terminale de la protéine (S1), dans la boucle reliant les hélices *α1* et *α2* ainsi que dans les portions d'hélices adjacentes (S2), et dans la boucle reliant *α3* et *α4* et l'extrémité N-terminale de l'hélice *α4* (S3). Dans le domaine de liaison à l'ADN, les quatre *IDSs* sont formés des résidus de la boucle reliant les brins *c* et *c'* (S4 et S5), de la boucle reliant les brins *c'* et *e* (S6) et de la boucle reliant les brins *e* et *f* (S7).

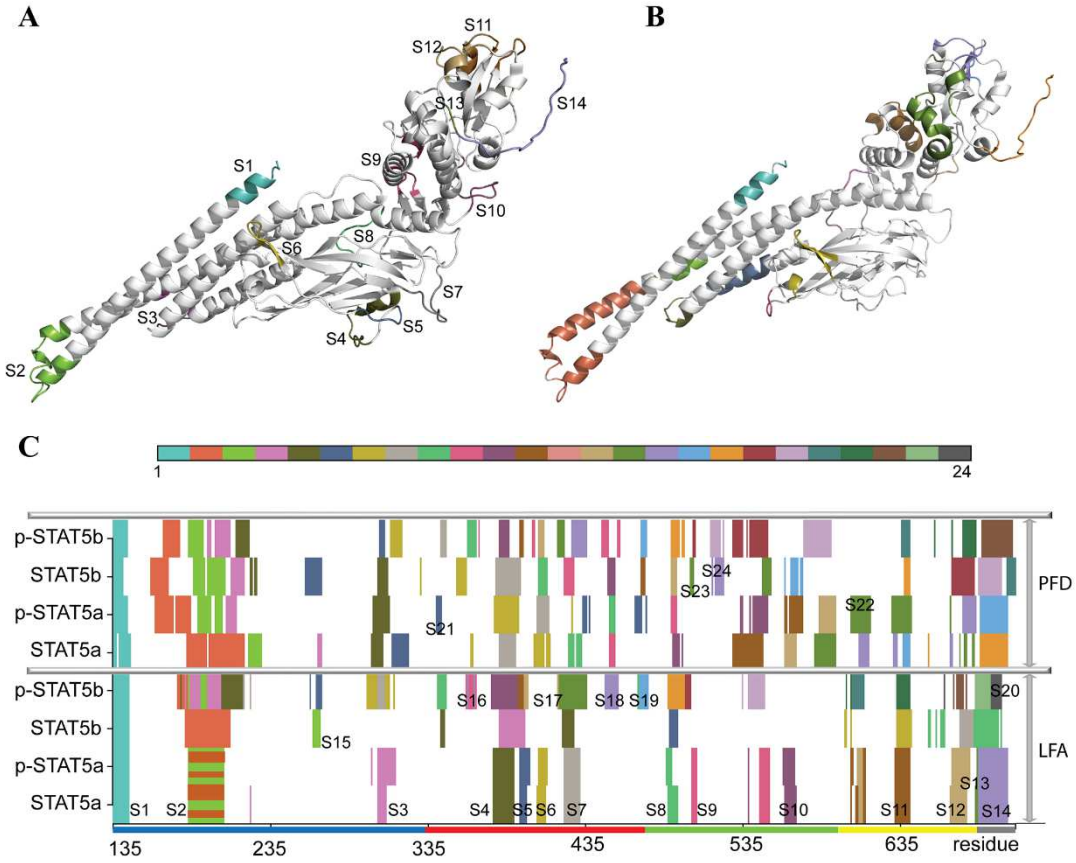


Figure 44: **Position des Segments Dynamiques Indépendants identifiés dans les protéines STAT5s.** (haut) Représentation tridimensionnelle de la position des *IDSs* identifié par les algorithmes LFA (**A**) et PFD (**B**). Les *IDSs* sont indiqués par la S_i, où i = 1…24. (bas, **C**) Représentation graphique de la position des *IDSs* pour chaque système, par les deux méthodes. Chaque couleur représente un *IDS*; les *IDSs* superposables sont hachurés.

Les trois *IDSs* dans le LD couvrent l'extrémité C-terminale de l'hélice *α5* et s'étendent sur la boucle à proximité (S8), la boucle entre le brin *h* et les hélices *α6*, *α7'* et la boucle reliant l'hélice *α7'* et le brin *i* (S9), la boucle reliant l'hélice *α7''* et *α8* (S10). Le domaine SH2 est couvert par deux *IDSs* : S11, qui implique les résidus de l'extrémité C-terminale de l'hélice *αA* et les boucles entre l'hélice *αA* et le brin *A*, et entre les brins *B* et *C*, et se superpose en partie avec S12, qui recouvre les résidus de la boucle entre l'hélice *αA* et le brin *A*, la fin de la boucle connectant les hélices *αC* et *αD* et l'extrémité N-terminale de l'hélice *αD*. Les deux derniers *IDSs*, S13 et S14, se recouvrent presque parfaitement, et recouvrent la queue (phospho-)tyrosyl. Dans STAT5b, presque tous les *IDSs* identifiés (S1-2, S4, S7-8 et S12-14), soit huit des neuf *IDSs*, correspondent à ceux observés

dans STAT5a (*cf.* Figure 44 C). Cinq *IDSs* de STAT5a ne sont pas retrouvés chez STAT5b (S3, S5-6, S9 et S10), alors qu'un court *IDSs* (S15, couvrant les résidus 263-267) a été identifié dans le CCD, au niveau de la boucle reliant les hélices *α2* et *α3*. D'autre part, S7 est étendu à la boucle reliant les brins *a'* et *b*. Concernant pSTAT5b, les *IDSs* sont localisés principalement aux mêmes endroits que pour STAT5a (S1-4, S8-9 et S11-14) ou STAT5b (S15), ou encore divisés en deux *IDSs* individuels (l'*IDSs* étendu S7 de STAT5b est divisé en S7 et S17 chez pSTAT5b). Trois nouveaux *IDSs* (S16, S18-19) ont également été identifiés dans le DBD, au niveau de la boucle entre les brins *b* et *c*, les brins *f* et *g*, et le brin *g'* à l'hélice *α5*, respectivement. La plupart des *IDSs* trouvés dans pSTAT5b se chevauchent.

Nous avons ensuite exploré les *IDSs* dans STAT5 en utilisant une nouvelle approche, appelée « Décomposition des Traits Principaux », et développée en collaboration avec le Pr. Alain Trouvé (CMLA, ENS Cachan – CNRS, UMR 8536), et plusieurs étudiants du département de Mathématiques de l'ENS Cachan. Cette méthode relocalise la majorité des *IDSs* identifiés par la méthode LFA, en particulier S1-4, S7-8 et S11-14 (*cf.* Figure 44 B et C). L'*IDS* S2 détecté par l'algorithme LFA au niveau du CCD distal est interprété comme deux (STAT5a), trois (pSTAT5a) ou quatre (STAT5b et pSTAT5b) *IDSs* discrets, positionnés soit sur la boucle ou sur les extrémités distales des hélices *α1* et/ou *α2*. L'application de l'algorithme PFD conduit à la prédiction d'*IDSs* bien définis et discrets dans toutes les protéines, de manière à ce que chaque résidu impliqué dans un *IDSs* soit rattaché uniquement à son meilleur prédicteur *(cf.* paragraphe IV.E du chapitre 2*)*.

La comparaison des résultats obtenus à l'aide de ces deux méthodes indépendantes (LFA et PFD) montre qu'elles permettent de décrire la dynamique locale des systèmes étudiés de manière similaire, ce qui valide la définition du concept des *IDS*s Cependant, l'algorithme PFD est plus rapide et fournit des informations supplémentaires, telles qu'une quantification de la variance cumulée de chaque *IDS*. L'analyse de la variance au cours des itérations de l'approche PFD indique que les mouvements associés aux *IDSs* constituent moins de 20% des mouvements de STAT5, alors que les mouvements globaux, caractérisant notamment les mouvements concertés du domaine CCD et de l'ensemble domaine SH2/queue (phospho-)tyrosyl, représente environ 80% des fluctuations atomiques. De manière remarquable, le premier prédicteur prédit très bien la variance après retrait des mouvements globaux.

Les résultats obtenus avec l'algorithme PFD corrèlent bien en termes de position des *IDSs* sur toutes les protéines STAT5 et, de manière similaire à l'approche LFA, suggèrent des caractéristiques communes de la dynamique locale des différents systèmes (*cf.* Figure 44 C, Figure 45 et Figure 46). La région du CCD distal et l'extrémité N-terminale de nos modèles, les boucles du DBD qui contactent l'ADN, les boucles connectant les brins B et C du domaine SH2 et la queue (phospho-)tyrosyl sont toutes impliquées entièrement ou partiellement dans des *IDSs* retrouvés dans toutes les isoformes, indépendamment du statut de phosphorylation. Néanmoins, des différences dans les propriétés des *IDSs* sont observées entre les formes de STAT5 et peuvent être reliées aux variations de séquence et/ou au statut de phosphorylation.

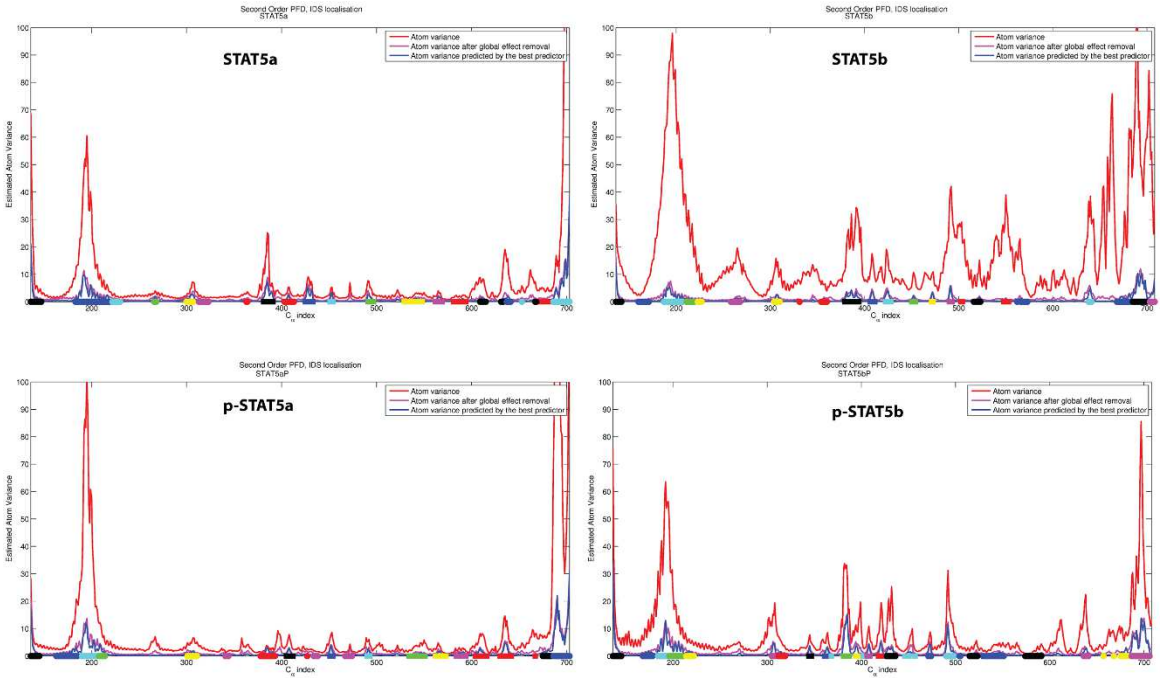


***Figure 45:* Réduction des variances atomiques au cours des itérations de l'algorithme PFD.** Les courbes en rouge indiquent les variances initiales, les courbes en magenta la variance du système après retrait des mouvements globaux (q = 6), les courbes en bleu donnent la variance expliquée par le meilleur prédicteur. Les *IDSs* sont indiqués le long de l'axe des abscisses.

Trois *IDSs* constituent des exemples représentatifs en lien avec les variations de la séquence peptidique de STAT5a et STAT5b : (i) S22 (*cf.* Figure 44 C), qui couvre la boucle connectant l'hélice αA au brin A (domaine SH2) n'est détecté que dans les systèmes STAT5a et pSTAT5a ; (ii) les *IDSs* S23 (couvrant la boucle suivant l'hélice α6) et (iii) S24 (qui couvre l'extrémité C-terminale de l'hélice α6 et le début de la boucle suivante) sont spécifiques des systèmes STAT5b et pSTAT5b. Ces *IDSs* peuvent être des extensions d'autres qui existent également dans les autres systèmes : S23 est ainsi le prolongement de S9. Lorsque l'on regarde les différences entre les formes phosphorylées et non-phosphorylées, 1 *IDS* présent au niveau de la boucle reliant les hélices α2 et α3 (S15) semble spécifique des formes non-phosphorylées alors que le brin a' est impliqué dans l'*IDS* S21 dans les deux formes phosphorylées, soit sous la forme d'un *IDS* individuel (pSTAT5b) ou étendu (pSTAT5a).

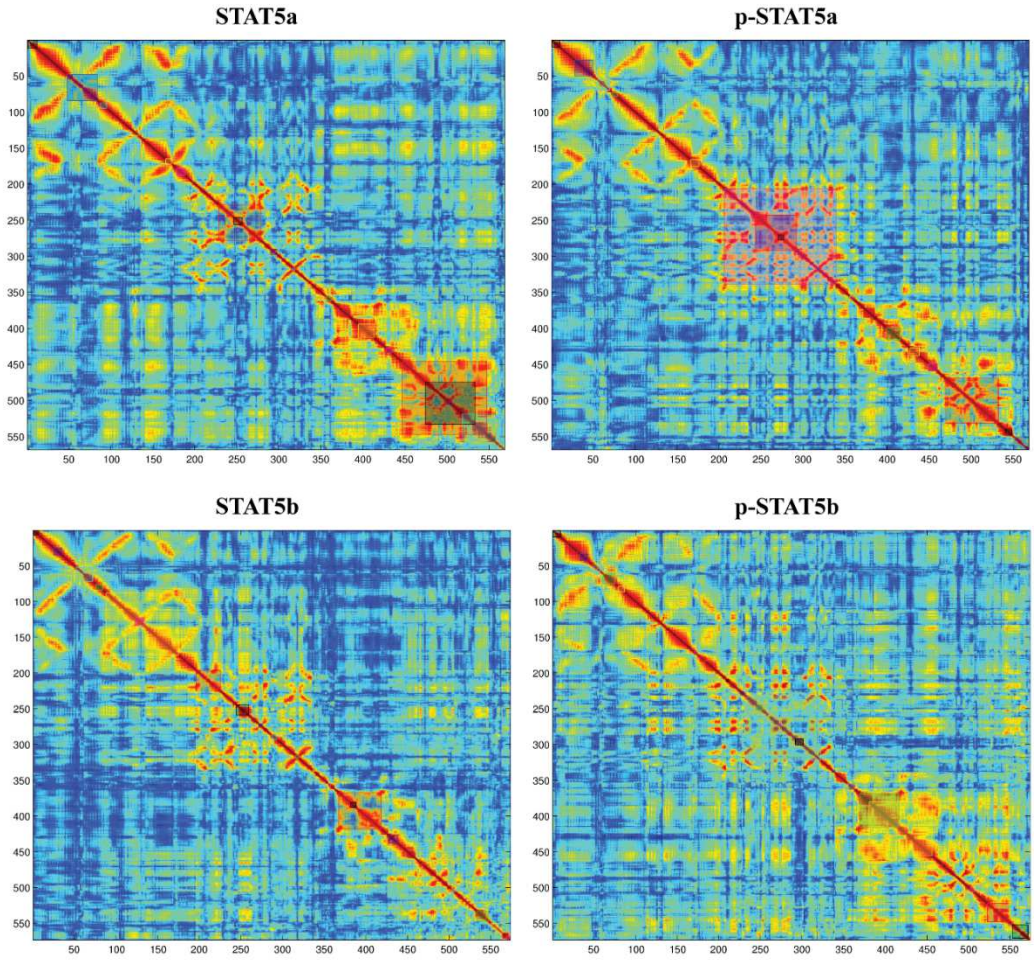


Figure 46: **Corrélations canoniques résiduelles obtenues après le retrait des mouvements globaux (q=6).** Les *IDSs* prédits par PFD sont indiqués par les rectangles.

Nous avons démontré un couplage qualitatif des mouvements protéiques entre les différents domaines dans les protéines STAT5. L'utilisation de deux algorithmes différents, LFA et PFD, nous a permis d'identifier les régions de STAT5 dont la dynamique interne présente les caractéristiques les plus frappantes. Les *IDSs* détectés par LFA sont comparables dans STAT5a et p-STAT5a, indiquant un modèle commun des dynamiques locales dans ces systèmes. En revanche, les *IDSs* dans STAT5b affichent une grande variabilité entre les espèces non phosphorylés et phosphorylés. Le CCD distal affiche un chevauchement parfait (STAT5a, p-STAT5a et STAT5b) ou partiel de l'*IDS* S2 (p-STAT5b), dénotant des mouvements locaux bien conservées, qui sont anti-corrélés avec les mouvements du CCD proximal adjacent. Le domaine SH2 affiche plusieurs chevauchent partiellement d'*IDSs*, indiquant une dynamique plus dissociée entre les différents éléments structurels de ce domaine. Néanmoins, les caractéristiques spécifiques des *IDSs* dans les différentes protéines STAT5 sont distinguées et peuvent être associées à leurs particularités liées à la séquence peptidique et/ou à leur statut de phosphorylation. Les IDS basés sur l'algorithme PFD affichent des résultats plus clairs, en termes de nombre d'*IDSs* et de leur composition/localisation. Dans le cadre méthodologique, l'approche PFD est une méthode rapide et élégante permettant d'identifier les *IDSs* selon la

variance moyenne normalisée des résidus. Ainsi, il fournit un outil analytique efficace et avantageux pour explorer la dynamique des protéines.

Fait intéressant, à proximité des *IDSs* spécifiques à une isoforme de STAT5 donnée ou un état de phosphorylation, aucune différence de séquence primaire (un remplacement de résidus de points ou une insertion) n'est trouvée, indiquant des effets dépendant de la séquence à longue portée sur la dynamique locale. La phosphorylation de la tyrosine spécifique peut induire un changement mineur dans la structure, mais modifie radicalement les fonctions des protéines, par exemple en produisant de nouveaux sites de liaison spécifiques. L'analyse des moyens de communication entre sites distants constitue ainsi un enjeu important de la description exhaustive de la mécanique des protéines.

2. Calcul des chemins de communication

Pour explorer le phénomène de communication entre des sites spatialement distants de STAT5, nous avons généré les *chemins de communication* (*Communication Pathways*, *CPs*) pour chaque modèle à l'aide de MONETA. Le paysage général des *CPs*, présenté sous forme de graphe 2D qui montre l'efficacité de la communication du réseau inter-résidus, et mis en évidence des différences dans les motifs, d'abord entre les isoformes a et b de STAT5, et ensuite entre les formes phosphorylées et non-phosphorylées.

Les séquences des deux isoformes STAT5a et STAT5b divergent par l'insertion de 5 résidus entre le domaine SH2 et la queue (phospho-)tyrosyl chez STAT5b, et par une série de remplacement ponctuels de résidus dans les différents domaines. Nous avons porté une attention particulière à ces résidus et à leur environnement. Presque toutes les régions des paysages de communication de ces isoformes montrent des différences significatives. Par exemple, le DBD contient 5 sites de remplacement polymorphes, et est caractérisé par un grand nombre de *CPs* entre les résidus de l'hélice *α5* et la boucle reliant les brins *e* et *f* dans STAT5a et pSTAT5a, alors qu'aucun *CP* similaire n'est trouvé dans STAT5b ou pSTAT5b (*cf.* Figure 47 A, région délimitée par le rond). Au niveau du domaine SH2 (*cf.* Figure 47 A, entourée en ovale), l'insertion du fragment CESAT dans l'isoforme STAT5b et la présence de trois sites différents montrent une différence de la communication entre les brins *B* et *C*. De nombreux *CPs* sont retrouvés entre les brins pour STAT5a et pSTAT5a, qui ne sont pas trouvés dans les protéines STAT5b. De manière intéressante, la plupart des chemins sont retrouvés entre des résidus conservés (I629 – W631 du brin *B* et W641 – N642 dans le brin *C*).

En comparant les formes phosphorylées aux formes non-phosphorylées de STAT5, une attention particulière a été portée aux chemins de communication transitant par le résidu Y694/699. Ce résidu tyrosine, crucial pour l'activité de STAT5, est directement engagé dans un *CP* seulement dans une protéine, STAT5a, et ne produit que des chemins à faible portée vers les résidus proches de la queue phosphotyrosyl. Dans les autres protéines, aucun *CP* n'a été identifié à partir du résidu Y694/699. Cependant, l'analyse de l'ensemble du réseau a établi un effet longue portée de la phosphorylation. En particulier, le CCD proximal, composé de la partie

proximale des hélices *α1-α2* et des hélices *α3-α4*, montre une densité de résidus fortement connectés dans les formes non-phosphorylées de STAT5, bien que distant d'environ 50Å du résidu Y694 – 699. Dans les deux STAT5 phosphorylées, cette région montre des résidus communiquant de manière plus modérée. Afin de quantifier ces différences, nous avons calculé (i) le nombre de *CP* entre chaque paire d'hélice et (ii) le nombre de résidus connectés par au moins un *CP* à un résidu d'une autre hélice (*cf.* Tableau 8). Une communication entre les hélices *α1* et *α2* est observée dans toutes les protéines, mais le nombre de *CPs* et le nombre de résidus connectés sont clairement augmentés dans les STAT5s non-phosphorylées. De manière similaire, la communication entre les hélices *α1* et *α3* est considérablement améliorée dans les formes non-phosphorylées de STAT5, alors que cette communication est faible (pSTAT5a) voir absente (pSTAT5b) dans les formes phosphorylées (*cf.* Tableau 8).

Tableau 8: Communication de la communication inter-résidus entre les hélices du CCD. Pour chaque paire d'hélice, les valeurs indiquent le nombre de résidus connectés par au moins un *CP*, et le nombre total de *CPs* (entre parenthèses).

Paire d'hélices	STAT5a	STAT5b	pSTAT5a	pSTAT5b
α1 – α2	71(1548)	58(1312)	8(26)	32(497)
α1 – α3	114(5807)	139(10450)	23(174)	0(0)
α1 – α4	0(0)	0(0)	1(1)	0(0)
α2 – α3	11(64)	15(182)	10(39)	2(3)
α2 – α4	11(34)	1(1)	24(148)	0(0)
α3 – α4	12(26)	0(0)	0(0)	0(0)
Total	219(7479)	213(13583)	66(388)	34(500)

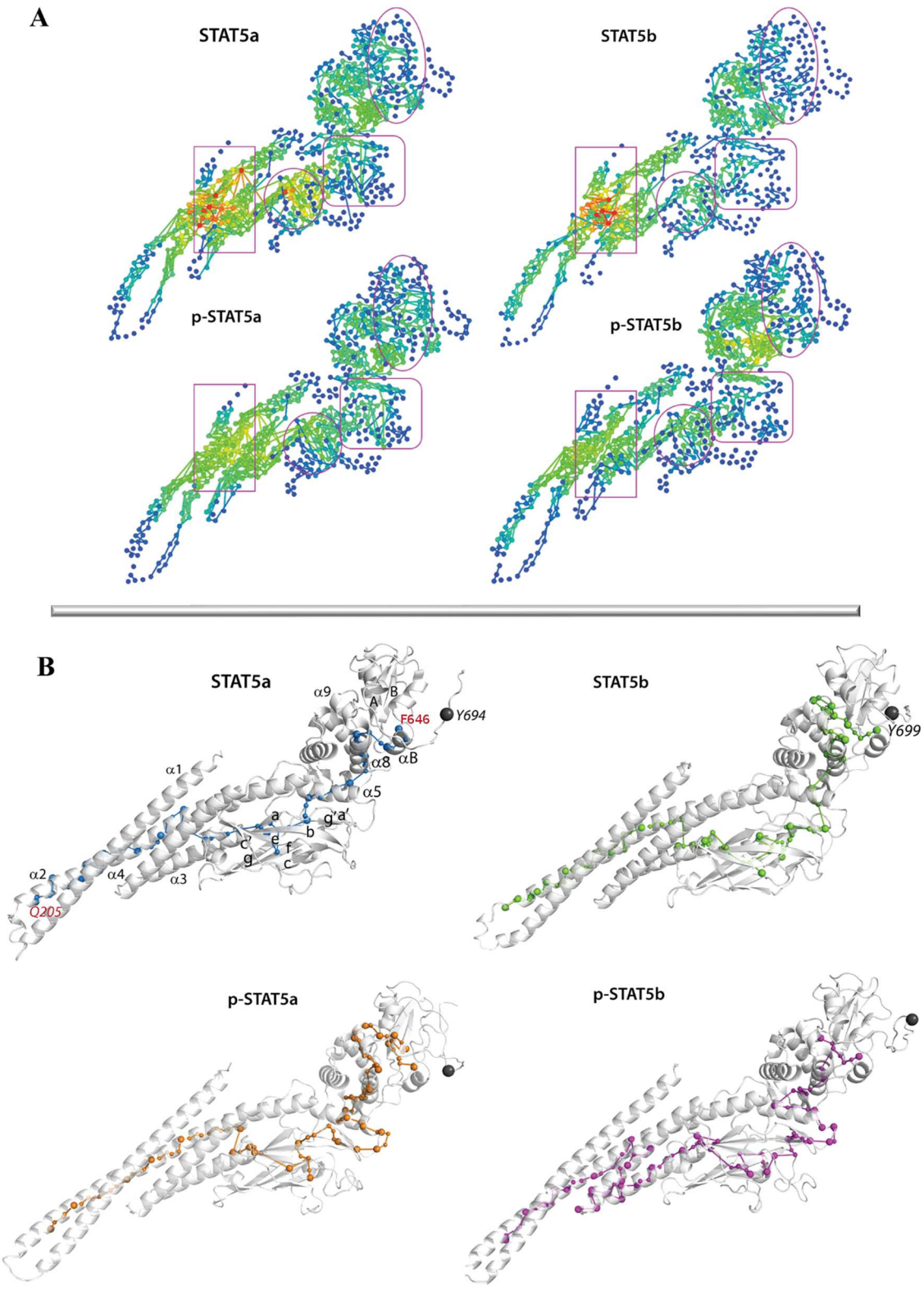


Figure 47 : **Réseau de communication de STAT5.** (**A**) Paysage général des chemins de communication inter-résidus représentés sous forme de réseau 2D. Chaque résidu est représenté par un nœud et les *CPs* sont dessinés par les lignes reliant les nœuds. Les résidus sont colorés en fonction de leur efficacité à communiquer (*Communication Efficiency, CE*), à savoir le nomre de chemins de communication qui passent par le résidu. Les résidus présentant une faible *CE* sont représentés en bleu, les résidus rouge ont une excellent *CE*. (**B**) Représentation 3D du pus court chemin entre les résidus Q205 et F646, sous forme de traits de couleur reliant les atomes des différents résidus traversés, représentés par des boules. Le carbone α du résidu Y694/699 est représenté oar une boule grise.

L'analyse détaillée des *CPs* a mis en évidence leur modification dans les domaines SH2, DBD et LD dans les formes phosphorylées comparativement aux protéines non-phosphorylées. Le chemin intramoléculaire le plus court (définit comme la succession de *CPs* impliquant le plus petit nombre de résidus pour relier 2 résidus distants) entre les résidus F646, situé à l'extrémité C-terminale du brin C dans le domaine SH2, et Q205, situé à l'extrémité N-terminale de l'hélice *α2* du CCD, est tracé à travers la structure 3D des protéines (*cf.* Figure 47 A**)**. Ce chemin générique connecte deux sites spatialement distants, Q205 et F646 étant séparés par une distance de plus de 100 Å. Le chemin Q205 - F646, qui transmet une information du CCD au SH2, est le plus court dans STAT5a, comparé à pSTAT5a, STAT5b et pSTAT5b. D'une manière plus générale, la longueur du chemin Q205 - F646 est supérieure dans les formes phosphorylées. Pour STAT5a, le chemin le plus court entre Q205 et F646 est composé de 18 *CPs*, comparé à 25 *CPs* pour pSTAT5a. Aucune succession de chemins de communication ne peut relier Q205 à F646, alors qu'un chemin de 30 *CPs* est observé chez STAT5b. L'interruption des chemins de communication de pSTAT5b se réalise à deux niveaux, entre le DBD et LD, et entre le CCD et le DBD. Ces effets pourraient refléter l'impact de la phosphorylation du résidu de tyrosine, qui perturbe localement la structure (queue phosphotyrosyl) mais également affecte des sites distants de la protéine.

Une caractéristique frappante de ce plus court chemin de communication Q205 - F646 entre sites distants est le circuit qu'il réalise dans le domaine SH2 et son passage du DBD vers le LD. Dans STAT5a, la communication *α8* – *αB* est directe, alors que dans STAT5b, les brins *A* et *B* ainsi que l'hélice *α7* sont impliqués (*cf.* Figure 47 B), faisant ainsi un itinéraire à travers tout le domaine SH2. Nous avons relevé la même observation dans pSTAT5a. Dans pSTAT5b, l'itinéraire n'est intermédiaire, entre les deux extrêmes décrits ci-dessus. Le passage du CCD au LD révèle également des différences importantes. Le passage implique un minimum de résidus dans STAT5a, alors que dans les deux formes de STAT5b, il est allongé de manière importante, et passe notamment pour tous les résidus le long de la boucle entre le brin *g´* et l'hélice *α5*.

F. Détection des poches à la surface des STAT5.

Un des éléments qui limite la recherche de nouvelles molécules inhibitrices en utilisant des connaissances structurales est la faible quantité de données relatives à la protéine cible. L'identification et la caractérisation des poches de liaison de petites molécules sont des facteurs cruciaux pour la recherche de « touches » (*hit*). Traditionnellement, la recherche de poches est effectuée sur les structures cristallographiques ou sur des modèles dits rigides. Les simulations de dynamique moléculaire peuvent être utiles dans la découverte de nouveaux sites de liaison, à travers l'exploration de milliers de conformations de protéines décrivant le comportement structurel et dynamique des macromolécules.

L'événement central dans la fonction STAT est une étape de dimérisation suivant la phosphorylation sur le résidu tyrosine spécifique 694 (STAT5a) ou 699 (STAT5b)[611]. Le ciblage du site de dimérisation représente donc un site de liaison potentiel pour de petites molécules qui

pourraient empêcher la liaison de la phosphotyrosine à son site cible. Une autre stratégie pourrait consister à inhiber les changements conformationnels des protéines STATs nécessaires pour le processus de dimérisation. Dans la forme dimérique parallèle de STAT3, les résidus K591, R609, S611 et S613, situés dans le domaine SH2, forment des interactions polaires directes avec la phosphotyrosine pY705[132], désignant ce site comme crucial pour les fonctions biologiques des protéines STATs. La surface des protéines STAT5 à proximité de ces résidus fonctionnellement cruciaux et de leurs propriétés a été analysée avec MDpocket[592].

Nous avons identifié deux poches adjacentes, P1 et P2, situés entre les domaines LD et SH2 (*cf.* Figure 48 A). La poche P1, délimitée par les hélices *α6* et *α7*, la boucle reliant ces hélices, et l'hélice *αA* du domaine SH2, a été retrouvée dans toutes les protéines simulées (*cf.* Figure 48 D). La deuxième poche, P2, trouvés dans le domaine SH2 entre l'hélice *αA* et le feuillet β *AB*, a également été systématiquement observée, sauf dans la deuxième réplique de STAT5a non phosphorylée (*cf.* Figure 48 C). Ces deux poches sont séparées par les résidus K600, R618, S620 et S622. Nous avons étudié davantage le niveau de conservation des résidus de cette région entre toutes les protéines STATs. Nous avons ainsi constaté que les résidus à proximité de P2 sont très bien conservés alors que plusieurs résidus délimitant P1 sont moins conservés dans la famille de protéine STAT (*cf.* Figure 48 B).

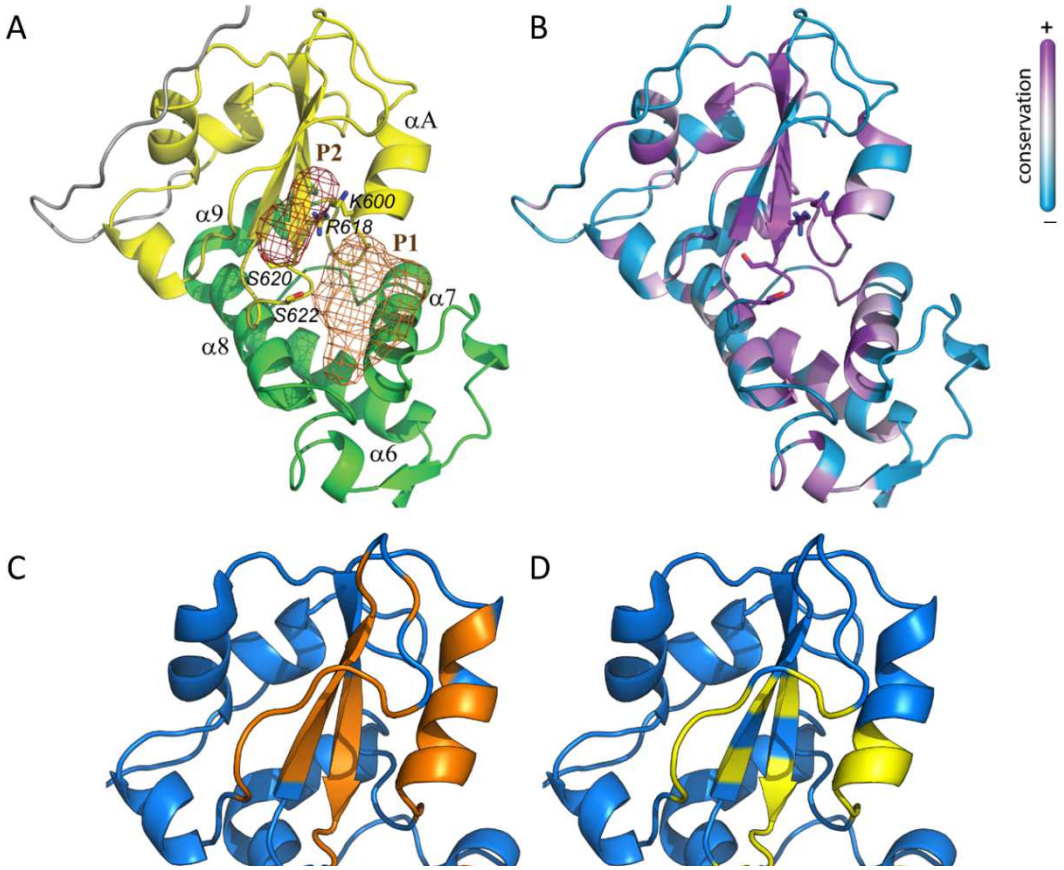


Figure 48: **Poches détectées à la surface de STAT5.** (**A**) Les deux poches PA et P2 (en brun et orange, respectivement) sont trouvées entre les domaines SH2 et LD. Les chaines latérales des résidus K700, R618, S620 et S622 sont représentées en bâtons. (**B**) Conservation des résidus au sein de la famille des protéines STATs. (**C**) Les résidus qui définissent la poche P2 sont représentés en orange. (**D**) Les résidus qui définissent la poche P1 sont représentés en jaune.

Afin de mieux caractériser les poches P1 et P2 et leur environnement, une deuxième analyse par MDpocket a été effectuée. Le volume mesuré pour les poches P1 et P2 sur les simulations dynamiques est représenté sur la Figure 49. Le volume des deux poches oscille au cours de la simulation, révélant le comportement dynamique des dimensions des poches. La comparaison des volumes des poches de STAT5a indique que P1 est grande (~ 500 $Å^3$) dans les deux formes pSTAT5a (en jaune) et STAT5a (en bleu), tandis que P2 est très peu volumineuse, fermé ou égal à zéro. Dans STAT5b, les profils poches varient différemment dans les deux répliques de dynamique moléculaire. La taille de P1 dans la forme non phosphorylée (en vert) varie de 0 à 200 $Å^3$ et de 100 à 400 Å3 pour la première et la deuxième réplique, respectivement, tandis que le volume de P2 fluctue dans la plage de 300 à 600 $Å^3$. Pour STAT5b phosphorylé (en magenta), la poche P1 est grande, de manière similaire à STAT5a, et ces variations sont similaires au cours des deux simulations, alors que la taille de P2 est proche de zéro dans la première simulation et varie de 300 à 550 $Å^3$ dans la seconde. Cette analyse met en évidence que deux poches adjacentes, P1 et P2, situés entre les domaines LD et SH2 de STAT5, présentent différents profils dans les simulations des protéines étudiées: systématiquement, P1 est volumineux et P2 petit chez STAT5a, tandis que dans STAT5b, leurs profils sont très divergents au cours des deux simulations de la même espèce.

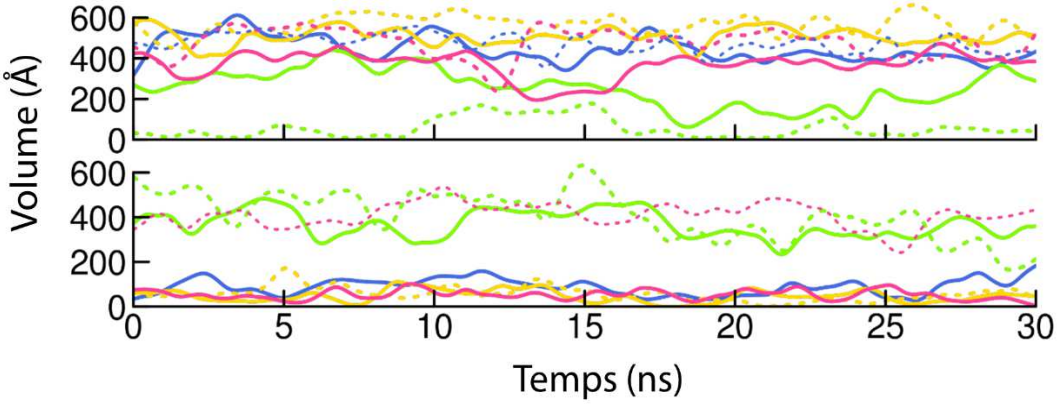


Figure 49: **Volumes de poches P1 (haut) et P2 (bas) au cours du temps.** Les deux simulations de STAT5a sont en bleu (trait plein et pointillé), pSTAT5a en jaune (trait plein et pointillé), STAT5b en vert (trait plein et pointillé) et pSTAT5b en magenta (trait plein et pointillé).

Dans STAT5, la phosphorylation de Y694/699 influe légèrement sur la structure des protéines et de la dynamique, tandis que la représentation modulaire de protéines obtenue avec MONETA produit explicitement des spécificités quantifiables pour chaque protéine de STAT5 non phosphorylée ou phosphorylée. Plusieurs caractéristiques uniques des STAT5s ont été mises en évidence à travers les paysages des chemins de communication, qui ont démontré que la liaison du groupement phosphate avec la tyrosine changeait considérablement les propriétés de la communication intra-protéine à longue distance. Pour obtenir une interprétation physiquement et fonctionnellement significative de nos résultats, nous avons superposé la plus courte voie de communication intramoléculaire reliant le domaine SH2 au domaine CCD (résidus Q205 - F646) dans chaque protéine, avec les poches localisées dans les domaines SH2-LD (*cf.* Figure 50). Cette représentation montre que les voies de communication

dans toutes les STAT5 étudiées sont localisées sur certains éléments structuraux communs (en particulier, les hélices *α7''*, *α8* et *α9*, et les brins *A* et *B*), qui constituent un *pipeline* moléculaire parfait pour la transmission de signaux entre des sites distants spatialement, séparés dans notre exemple par une distance supérieure à 100 Å. Nous avons constaté que le résidu F646 a été rapporté comme un point de mutation ayant un impact clinique sur la fonction de STAT5. En effet, la substitution naturelle du résidu F646 en sérine, la seconde mutation humaine rapportée, est associée à une déficience sévère en l'IGF-I, un dysfonctionnement immunitaire sévère sans aggravation de la fonction pulmonaire[612]. Par ailleurs, le résidu R618 qui participe au plus court chemin de communication intramoléculaire dans deux protéines, pSTAT5a et pSTAT5b, est conservé dans toutes les STATs, et se situe à la marge des deux cavités P1 et P2. Ces résidus sont identifiés comme primordiaux pour la fonction biologique des STATs[132]. Nous observons donc une co-localisation (i) de résidus importants pour la fonction de STAT5 et (ii) de poches de liaison potentielles.

Cette observation suggère une utilisation plausible des poches de STAT5 pour développer des inhibiteurs capables de moduler les propriétés de communication de cette protéine de signalisation. Cette modulation, inspirée par l'analyse des *CPs* et ciblant les voies de communication intramoléculaire, peut potentiellement bloquer plusieurs processus, comme la dimérisation, la liaison de l'ADN ou la reconnaissance de STAT5 par des activateurs. Les inhibiteurs connus, ciblant STAT5[421,423,425,428,613–615] sont actifs à des concentrations élevées qui ne sont pas applicables en clinique. Pour concevoir des inhibiteurs STAT5 de manière rationnelle dans le but de les rendre plus sélectifs, nous proposons d'explorer les poches de liaison décrites afin de perturber le réseau des voies de communication. Comme les motifs de communication ne sont pas conservés dans les protéines STAT5, concevoir des inhibiteurs de STAT5 sélectif envers une isoforme donnée est d'un grand intérêt. Établir *in vitro* et *in cellulo* les caractéristiques spécifiques de la dynamique conformationnelle et de la communication décrite dans les différents STAT5 aidera également à élucider leur rôle dans la signalisation cellulaire.

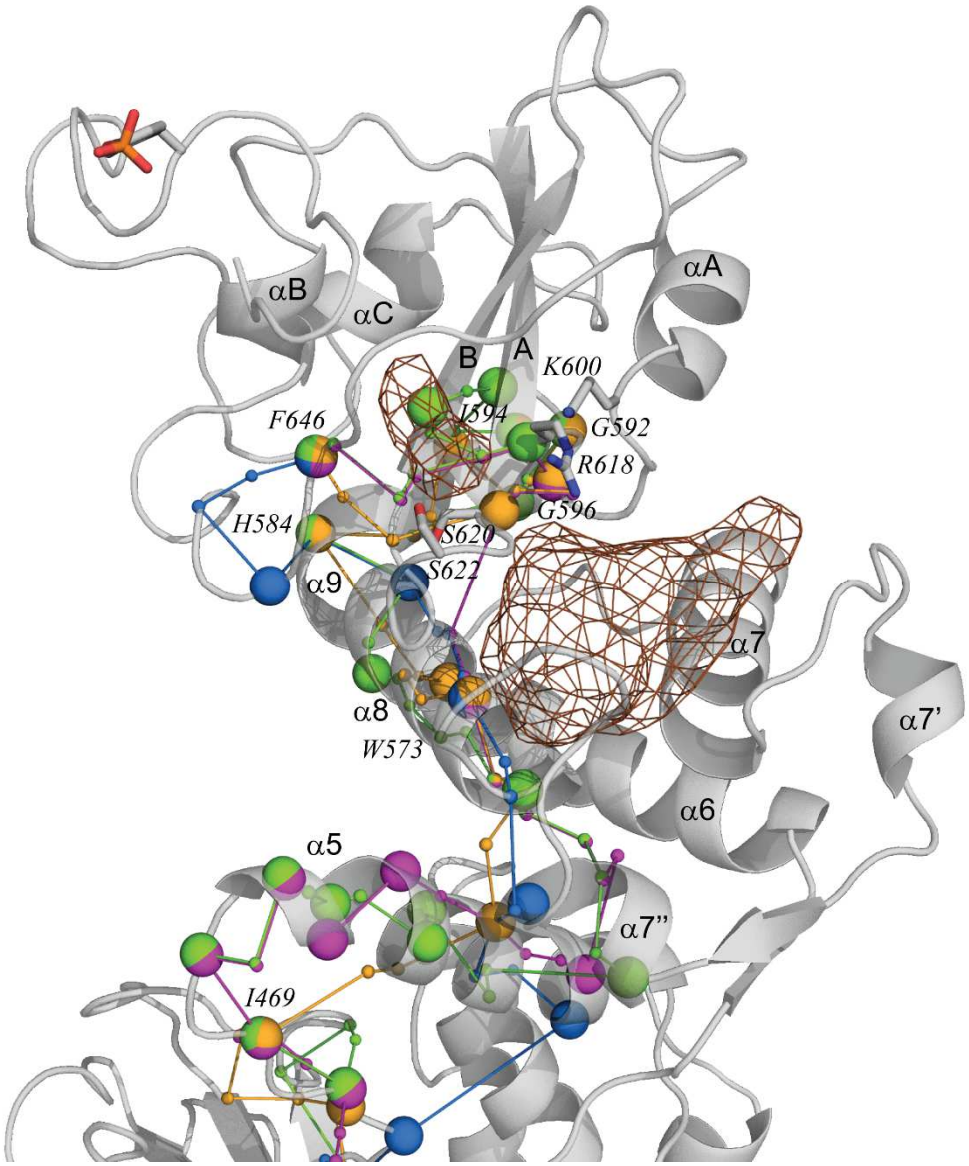


Figure 50: **Chemins de communication et poches de liaison dans les protéines STAT5.** Les plus courts chemins intramoléculaires entre Q205 et F646 pour chaque système sont superposés et représentés avec les poches P1 et P2 de pSTA5a. Les chemins de communication sont représentés par des traits reliant les boules, qui représentent les carbones α (grosses boules) ou un atome de la chaîne latérale (petite boule). Le code couleur représente les différentes protéines STAT5 : bleu, STAT5a ; jaune, pSTAT5a ; vert, STAT5b et magenta, pSTAT5b. Les boules multicolores sont des résidus empruntés par plusieurs chemins de communication.

II. Caractérisation des dimères de STAT5

A. Structures et organisation générale

Structures des modèles générés par homologie

Les modèles des dimères du *Core Fragment* (*CF*) de STAT5 ont été générés en utilisant les mêmes structures supports que pour la modélisation des protéines monomériques. Deux différents modèles ont été générés, dénommés dSTAT5a et dSTAT5b suivant qu'ils représentent un homodimère de STAT5a ou de STAT5b, respectivement. Chaque modèle comprend également une molécule d'ADN constituée de deux brins d'ADN de 18 bases, qui forment un double-brin de 17 paires de bases. Les nucléotides situés en 5' de chaque brin sont libres, *i.e.* non-engagés dans la formation d'une paire de bases. À la différence des protéines monomériques, toutes les espèces dimériques de STAT5 sont phosphorylées au niveau du résidu de tyrosine 694 (STAT5a) ou 699 (STAT5b), afin de reproduire l'arrangement dimérique parallèle lié à l'ADN (*cf.* paragraphe I.A.6 du chapitre 1). Pour générer les modèles de dimères de STAT5, nous avons utilisé la structure murine de STAT5a 1Y1U[118] et la structure murine de STAT3β liée à l'ADN 1BG1[132], ainsi que le même alignement de séquence entre les séquences des structures supports et la séquence cible (STAT5a et STAT5b).

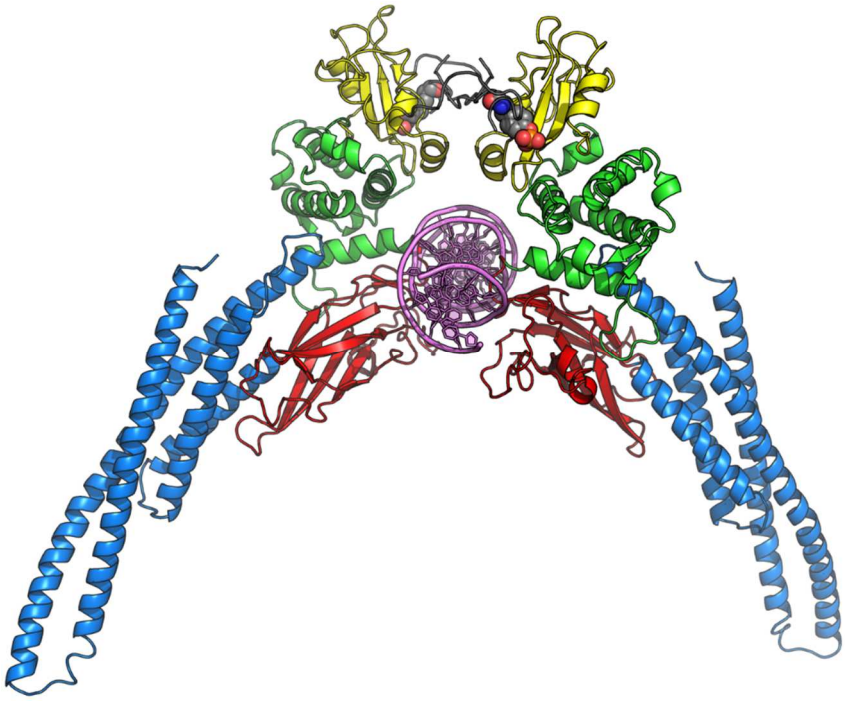


Figure 51: **Représentation du modèle généré par homologie de dSTAT5b.** Les domaines CCD sont colorés en bleu, les DBD sont en rouge, les LD en vert, les SH2 en jaune et la queue phosphotyrosyl en gris. Le double-brin d'ADN est coloré en violet. Les deux résidus de phosphotyrosine sont représentés en sphères.

L'analyse des données expérimentales concernant l'affinité de STAT5 pour différents motifs d'ADN a montré que STAT5 se lie à des motifs GAS (*Interferon-gamma activated sequence*). Ces motifs sont des séquences palindromiques, reconnues par les STATs. Nous avons donc modélisé la séquence pour laquelle STAT5 présente la meilleure affinité[530] à partir de la séquence GAS contenue dans la structure cristallographique 1BG1[132] en utilisant le logiciel *Coot*[531].

L'architecture générale des monomères impliqués dans la formation d'un dimère correspond à celui des espèces monomériques seules (*cf.* Figure 51). Ils sont constitués d'un domaine CCD formé de quatre hélices *α1* – *α4*, dont les deux premières (*α1* – *α2*) sont sensiblement plus longues et présentent plusieurs courbures. Dans les modèles dimériques, la longueur des hélices du CCD est très similaires à celles observées dans les espèces monomériques : l'hélice *α1* varie de 48 à 50 résidus, l'hélice *α2* est longue de 69 résidus, *α3* est constituée de 30 résidus et l'hélice *α4* de 23 résidus. Les domaines de liaison à l'ADN (DBD) sont également similaires, et présentent les 3 feuillets β constitués respectivement des brins *abe*, *cfg* et *a'g'*. D'autres brins courts sont également observés sur certains monomères. Par exemple, un brin *c'* est observé dans la boucle reliant les brins *c* et *e* du modèle de STAT5b, dans le prolongement du feuillet *cfg*. Un brin *x* est également observé dans les deux monomères de dSTAT5a et dSTAT5b, dans le prolongement du feuillet *abe*, dans la boucle entre les brins *c'* et *e*. Une ou plusieurs hélices α ou 3_{10} sont présentes de manière non constante, quelque résidus après la fin du brin *c*. La boucle reliant les brins *e et* f s'accommode systématiquement le long du double brin d'ADN et présente une surface d'interaction STAT5 – ADN importante. La partie de l'ADN contactée par STAT5 n'est en revanche que la chaîne principale et non les bases azotées, et ne semble donc pas le facteur déterminant de la spécificité de la reconnaissance STAT5 – ADN. De même, la boucle reliant les brins *a'* et *b* contacte le squelette de l'ADN, sans que les chaînes latérales des résidus n'entrent en contact avec les bases azotées. Ces données corrèlent bien avec les données expérimentales qui situent le site porteur de la spécificité au niveau du résidu H471[132,133], sur la boucle entre le brin *g'* et l'hélice *α5*.

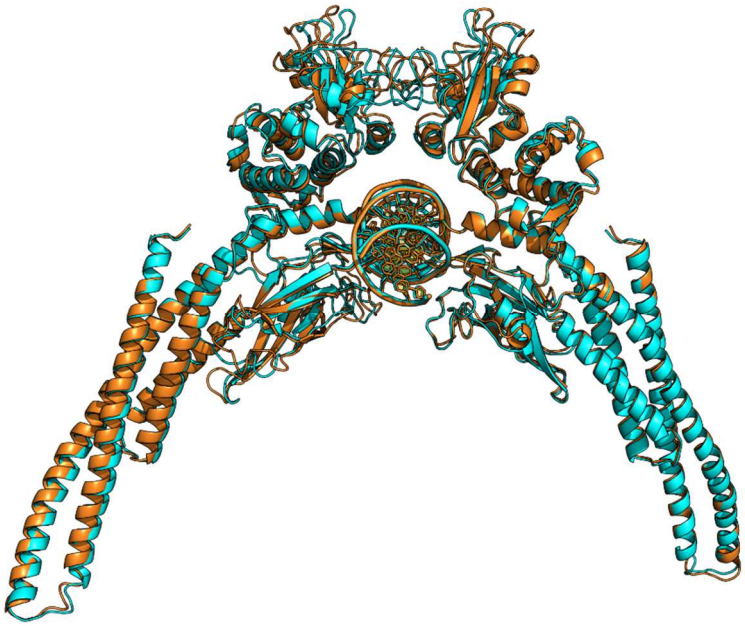


Figure 52: **Superposition des modèles dSTAT5a et dSTAT5b.** dSTAT5a est représenté en cyan alors que dSTAT5b est en orange.

La boucle *g′* – *a5* est systématiquement située dans un sillon large de l'ADN, pour tous les modèles générés (*cf.* Figure 53). De manière intéressante, alors que nous nous attendions à retrouver des interactions entre le résidu H471 et les motifs d'ADN reconnus spécifiquement, ces résidus ne sont pas tournés dans la direction des motifs TTC-GAA complémentaires (*cf.* Figure 53). Modeller[532], le logiciel que nous avons employé pour l'étape de modélisation par homologie, n'est pas capable de modéliser les molécules d'ADN. Plus précisément, modeller traite ce type de molécule de manière rigide et ne peut donc pas assurer l'optimisation précise des interfaces protéines – ADN. Afin de décrire plus précisément ces interfaces, nous avons utilisé le logiciel « *Protein-DNA Modeling Interface* »[616]. Ce logiciel génère puis évalue l'affinité d'une interface pour une séquence d'ADN donnée afin de prédire la conformation des chaînes latérales qui contactent l'ADN. Ainsi, les auteurs ont réussi à reproduire les résultats expérimentaux pour les doigts de zinc C_2H_2 [616]. Cette la méthode a permis de reproduire la spécificité de reconnaissance de motifs d'ADN d'un facteur de transcription à partir des 21 structures connues de protéines homologues contenant 93 interfaces protéine – ADN. Nous avons appliqué l'algorithme en modifiant les séquences d'ADN afin d'évaluer l'affinité de STAT5 pour les différents motifs d'ADN, bien que cette approche ne permette pas de distinguer les états de protonation de l'histidine. Pour chaque site (*i.e.* monomère), les résidus de la boucle reliant les brins *a′* et *b*, la partie C-terminale de la boucle reliant les brins *e* et *f* ainsi que la boucle connectant le brin *g′* à l'hélice *a5* et les motifs TTC-GAA ont été considérés comme semi-flexibles : le logiciel a donc la possibilité de modifier la chaîne latérale de ces résidus, ainsi que les bases nucléotidiques. Les résultats obtenus n'ont malheureusement pas permis de différencier *in silico* les séquences d'ADN et donc de reproduire les résultats expérimentaux[530].

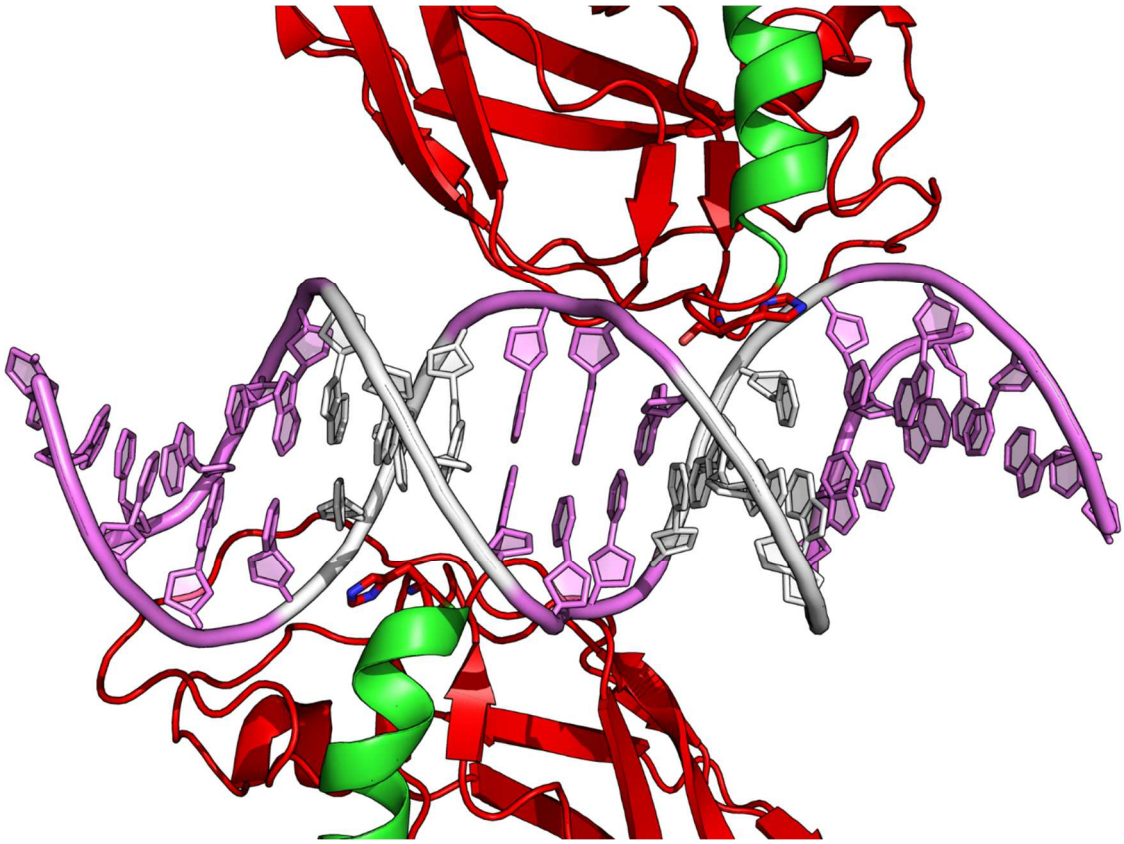


Figure 53: **Interface STAT5 / ADN des modèles par homologie.** Zoom sur l'interface entre le double brin d'ADN et dSTAT5b. Le DBD est en rouge, le LD en vert et l'ADN en violet. Les atomes lourds des résidus H471 des deux monomères sont représentés en bâton. La queue phosphotyrosyl, le domaine SH2 et le LD à l'exception de l'hélice *α5* ne sont pas représentés pour des raisons de clarté. Le motif de reconnaissance de l'ADN TTC-GAA est représenté en blanc.

D'autre part, nous avons constaté en initiant l'équilibration des systèmes en parallèle du travail réalisé avec le logiciel « *Protein-DNA Modeling Interface* » que les résidus à l'interface STAT5 – ADN sont capables de s'adapter à la surface du double brin d'ADN, et d'établir des liens directs avec certaines bases azotées. Nous supposons que des interfaces s'adapteront de manière optimale au cours de l'équilibration des systèmes avant les simulations de dynamique moléculaire de production.

Le domaine LD des dimères est très similaire aux modèles des espèces monomériques, et reste composé de 6 à 7 hélices, l'hélice *α7'* n'étant parfois pas détectée par l'algorithme DSSP, et du même court feuillet β antiparallèle. Le domaine SH2 est lui aussi semblable, et comprend un feuillet β composé des deux brins *A*, *B*, et un brin *C* qui n'est retrouvé que dans un monomère du modèle dSTAT5b. Ainsi, à l'inverse des espèces monomériques de STAT5 non-phosphorylées, toutes les formes phosphorylées, monomériques et dimériques, ne possèdent pas systématiquement un feuillet à 3 brins (*cf.* paragraphe I.A.2 du chapitre 3). De par leurs positions dans l'espace, les résidus de l'hélice *αB* peuvent contacter les atomes de la chaîne principale de l'ADN et former des contacts non spécifiques. La queue phosphotyrosyl dans les dimères présente un arrangement très différent de celui observé chez les formes monomériques : le résidu de phosphotyrosine est systématiquement lié au domaine SH2 lui faisant face et assure

ainsi la formation du dimère. De par ces contraintes supplémentaires, la position de la queue est semblable dans toutes les espèces dimériques alors que de grandes différences étaient observées dans les formes monomériques.

Tableau 9 : **Distance (en Å) des monomères individuels de dSTAT5 par rapport aux modèles monomériques.** La distance est mesurée à partir des carbones α uniquement.

	dSTAT5a		dSTAT5b	
STAT5a	0,713	0,692	0,784	0,515
pSTAT5a	1,201	1,249	1,025	1,098
STAT5b	0,686	0,882	1,266	0,646
pSTAT5b	1,382	1,415	1,456	1,359

La distance (évaluée par le RMSD) entre les modèles de STAT5 monomériques et les monomères issus des modèles dimériques est du même ordre de grandeur que la distance entre les modèles monomériques (*cf.* Tableau 9 et Tableau 6). Le RMSD entre les carbones α des deux modèles dimériques dSTAT5a et dSTAT5b est lui aussi du même ordre, à 0,996 Å, et reflète la grande similarité des deux modèles. Enfin, les RMSDs mesurés par rapport aux structures supports, 1Y1U et 1BG1, sont également similaires à ceux observés pour les modèles de STAT5 monomérique, des valeurs deux fois plus faibles étant mesurées pour le support 1Y1U que pour 1BG1. Ces valeurs semblent logiques au vu de la plus grande identité de séquence entre les formes humaines et murines de STAT5, qu'entre la forme humaine de STAT5 et la forme murine de STAT3 (*cf.* Tableau 10).

Tableau 10 : **Distances des modèles dimériques par rapport aux structures supports.**

	dSTAT5a		dSTAT5b	
1Y1U	0,703	0 ,698	0,801	0,460
1BG1	2 ,387	1,811	1,137	1,578

Analyse des profils de déviations au cours de la trajectoire :

Les profils de déviation ont été calculés à la fois par rapport à la conformation initiale de la simulation et par rapport à la structure moyenne (*cf.* Figure 54). Les simulations des espèces dSTAT5^{HIP} (*cf.* Figure 54) sont toutes les deux stables, et ne présentent que des variations faibles (< 5Å) (*cf.* Figure 54). Pour les autres espèces (STAT5^{HID} et STAT5^{HIE}), on remarque des différences parfois importantes entre les simulations, avec notamment de larges variations de RMSD pour les espèces dimériques de STAT5a^{HIE} et STAT5b^{HIE} comparées à la conformation initiale, variations dont la valeur maximale reste inférieure à 11Å. Ces variations augmentent rapidement entre 0 et 20 nanosecondes de simulation, mais qu'elles décroissent légèrement par la suite jusqu'à des valeurs de 6 – 7 Å (*cf.* Figure 54, courbes violette, à gauche). Lorsqu'on compare ces variations à la conformation moyenne, on observe que les déviations sont bien plus faibles, de l'ordre de 6 Å au maximum (*cf.* Figure 54, courbes violettes, à droite). Dans le cas

de STAT5^{HID}, les variations observées sont également faibles à l'exception de STAT5b^{HIP} qui montre une augmentation soudaine des valeurs de RMSD aux alentours de 25 nanosecondes de simulation, quelle que soit la structure de référence (la conformation à t = 0 ns ou la conformation moyenne) utilisée pour les calculs. Ces résultats semblent indiquer des mouvements amples de la dynamique globale des protéines STATs au cours des simulations. Cependant, au vu de la forme générale du dimère de STAT5 lié à l'ADN (*cf.* Figure 51), il est intéressant de déterminer si les variations du RMSD sont dues à un réarrangement structural de certains domaines, ou si des régions de la protéine présentent une dynamique locale au cours des simulations, à l'image du CCD et de la queue (phospho-)tyrosyl dans les formes monomériques.

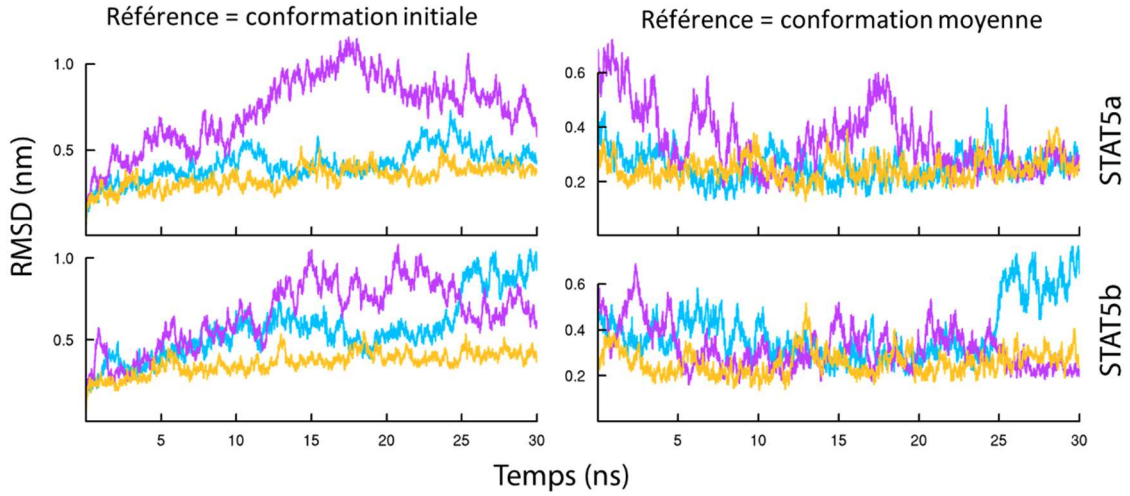


Figure 54 : **Profils de déviations des simulations de dynamique moléculaire des dimères dSTAT5/ADN.** La structure de référence est soit la structure initiale (gauche), soit la conformation moyenne de la trajectoire (à droite). Les simulations des dimères de STAT5a sont dans les cadrans supérieurs alors que les simulations des dimères de STAT5b sont dans les cadrans inférieurs. Les espèces dont le résidu H471 est protoné en δ sont en bleu ciel, en ε en violet et protoné sur les deux sites en orange.

Pour répondre à cette question, nous avons calculé les RMSDs de chaque domaine après superposition sur la structure initiale ou moyenne. Les domaines considérés sont les cinq domaines constituant le *Core Fragment* de STAT5 (CCD, DBD, LD, SH2 et queue phosphotyrosyl), ainsi que les brins d'ADN. Pour chaque conformation issue des simulations de dynamique moléculaire, nous avons superposé chacun de ces domaines sur sa position initiale à t = 0 ns ou sa position moyenne (*cf.* Figure 55). Nous pouvons observer que les domaines DBD, LD et SH2 montrent une grande stabilité, caractérisée par des valeurs de RMSD souvent inférieures à 0,2 nm et toujours inférieures à 0,3 nm. Le plateau de ces courbes est plus bas dans le cas de la comparaison à la conformation moyenne, dénotant une stabilisation rapide de ces domaines.

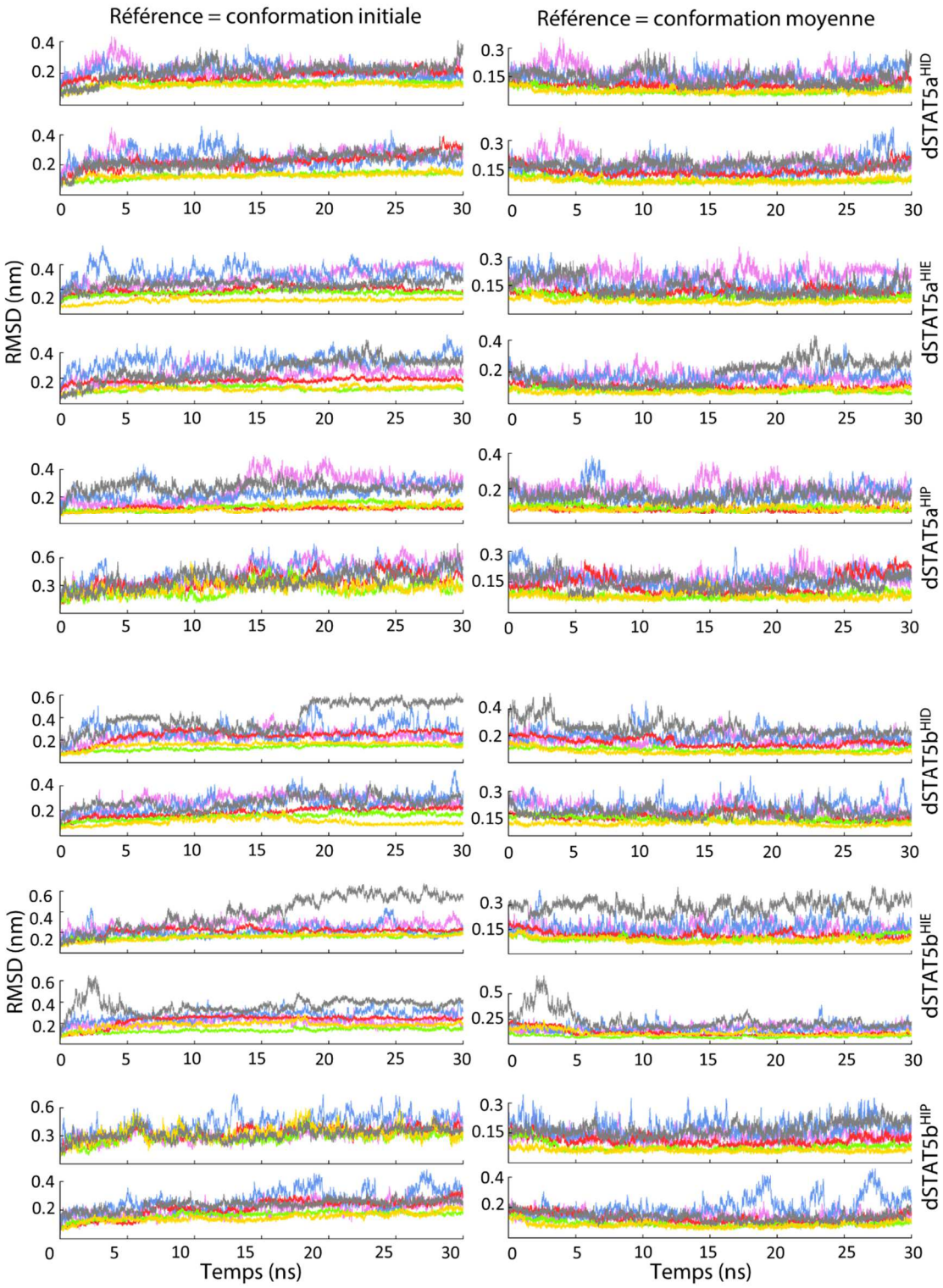


Figure 55 : **Profils de déviation des domaines de STAT5 au cours des simulations de dynamique moléculaire de dSTAT5/ADN.** Les RMSDs sont calculés en utilisant la conformation initiale (t = 0 ns, à gauche) ou la conformation moyenne (à droite). Pour chaque simulation, les deux monomères sont présentés sur un graphique séparé pour des raisons de clarté. Le CCD est en bleu, le DBD en rouge, le LD en vert, le SH2 en jaune, la queue phospho-tyrosyl en gris et l'ADN en rose.

Le CCD montre des déviations plus importantes, mais toujours inférieures à 0,5 Å. Plus intéressant, les variations peuvent être brusques, et les courbes sont moins stables que celles

correspondant aux autres domaines (*cf.* Figure 55, courbes bleues). Ces résultats indiquent que le CCD présente une dynamique interne, qui se poursuit au cours de la trajectoire, de manière similaire aux systèmes monomériques. Afin de déterminer quelles sont les zones du CCD qui ont les plus grands déplacements, nous avons superposé la conformation de STAT5b^{HIP} prise à 27,31 ns de simulation à la structure moyenne de cette dynamique. Cette conformation correspond à la plus haute variation de RMSD (4,55 Å) pour cette simulation comparée à la structure moyenne (*cf.* Figure 55, courbe bleue du graphique inférieur droit). Nous pouvons clairement voir sur la Figure 56 que le CCD distal s'est déplacé alors que le reste du CCD (extrémités proximales des hélices α1 et α2 ainsi que les hélices α3 et α4) ne s'est pas déplacé, ou très peu.

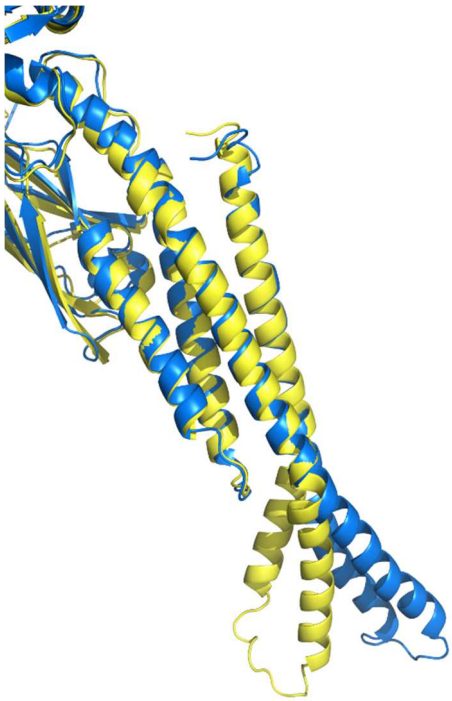


***Figure 56 :* Analyse de la flexibilité du CCD de dSTAT5b^{HIP}.** Déviation maximale du CCD au cours de la trajectoire. La structure moyenne est colorée en jaune, la conformation à 27,310 ns est en bleu.

Le plateau des courbes associées à la queue phosphotyrosyl est compris entre des valeurs de 0,2 à 0,6 Å (*cf.* Figure 55, courbes grises). Les fluctuations des courbes semblent indiquer l'existence de plusieurs arrangements différents qui sont parcourus au cours de la trajectoire (*cf.* Figure 55, dSTAT5b^{HID}). Ce phénomène est mis en avant par le fait que les variations des RMSDs calculés par rapport aux conformations moyennes sont plus stables que les courbes présentant les RMSDs calculés par rapport aux conformations initiales qui peuvent présenter des variations brusques. Par exemple, la transition d'une configuration à une autre de la queue phosphotyrosyl est illustrée par le bond du RMSD d'un monomère de dSTAT5b^{HID} (*cf.* Figure 57 et Figure 55). Au cours de la simulation de dynamique moléculaire de ce dimère, on peut observer un large déplacement (plus de 18 Å pour le résidu A690 entre la conformation initiale et la conformation finale) de la partie de la queue connectant le domaine SH2 au résidu phosphotyrosine (*cf.* Figure 57). La partie C-terminale de cette région est à l'interface entre les deux monomères et présente au contraire une dynamique très stable.

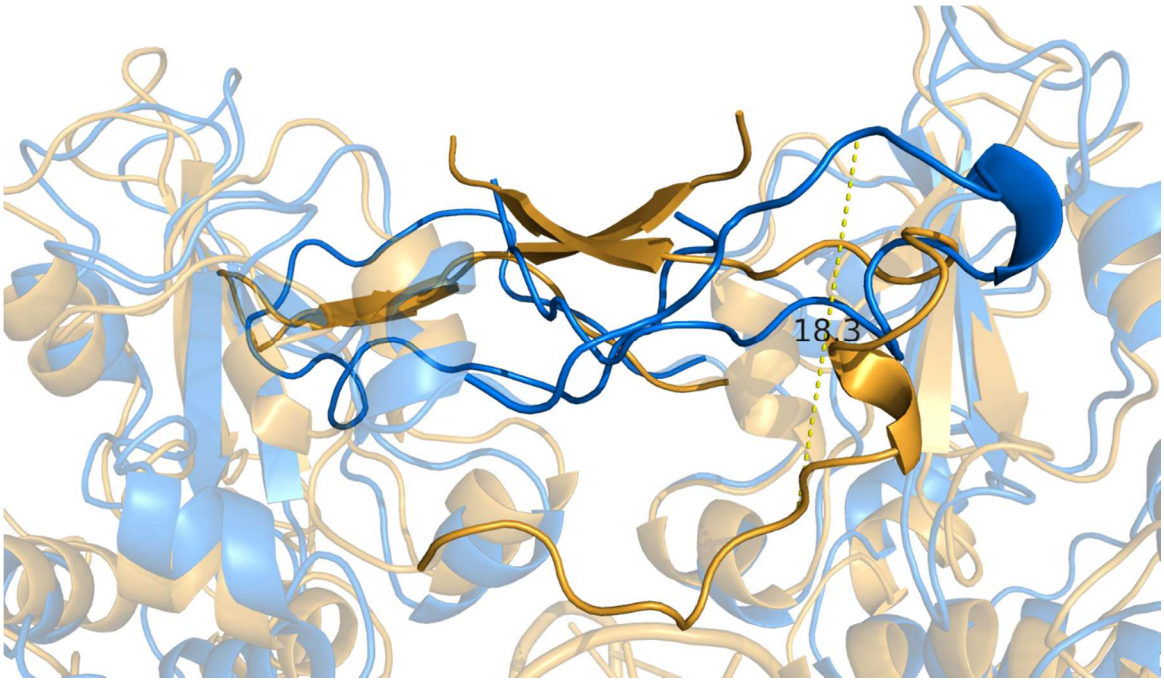


Figure 57 : **Déplacement de la queue phosphotyrosyl de STAT5b^{HID} au cours de la simulation de dynamique moléculaire.** La conformation initiale est en bleu, la conformation finale est en orange. La distance entre le carbone α du résidu 690 entre les deux conformations est indiquée par la ligne pointillée jaune. Pour des raisons de clarté, les domaines SH2, LD et l'ADN sont montrés en transparence.

Enfin, les courbes de RMSDs correspondant aux brins d'ADN montrent un plateau autour desquelles elles fluctuent (*cf.* Figure 55, courbes roses). Par comparaison aux domaines spatialement proches de STAT5 (DBD, LD et SH2, présentés par les courbes rouges, vertes et jaunes respectivement, Figure 55), l'ADN présente des fluctuations plus importantes. Chaque brin ne comportant que 18 résidus, les RMSD sont calculés en utilisant la position des 17 atomes de phosphore de la chaine principale. Puis, nous avons calculé le RMSD en utilisant les 18 atomes C1' et les 18 atomes C2 de chaque nucléotide. Les résultats obtenus sont similaires, ce qui démontre un plus grand déplacement des brins d'ADN comparativement aux domaines protéiques proches. Étant donné que les deux extrémités de chaque brin d'ADN est libre, les variations des courbes de RMSD pourraient correspondre aux déplacements des résidus nucléotidiques placés aux extrémités.

Afin de localiser plus finement les régions des dimères qui fluctuent le plus, nous avons calculé pour chaque simulation les RMSFs de chaque monomère composant les dimères de STAT5 et de chaque brin d'ADN. Nous retrouvons plusieurs similitudes avec les espèces monomériques de STAT5. La partie N-terminale de chaque monomère de dSTAT5 présente un pic important qui diminue très rapidement, en lien avec la structuration de l'hélice α1. Un second pic majeur est observé pour le CCD distal, à l'image des monomères. De nouveau, les hélices α1 – 4 de ce domaine affichent des fluctuations faibles et stables, alors que le CCD distal est très mobile. Les fluctuations atomiques de cette zone sont du même ordre que celles observées pour les espèces monomériques STAT5 libres, suggérant la présence de la même

organisation du CCD en deux modules : un module stable constitué de la partie proximale de α1 et α2 ainsi que de α3 et α4, ainsi qu'un module mobile (CCD distal) qui bouge autour du module fixe. Ce comportement est particulièrement bien illustré dans la Figure 56, et explique les variations de RMSD observées. Le déplacement de 'va – et – vient' du module mobile autour du module fixe entraîne l'oscillation de la courbe des RMSDs. Enfin, une différence notable peut être notée au niveau de la queue phosphotyrosyl. Un pic des RMSFs est systématiquement observé, mais l'extrémité C-terminale de la protéine présente invariablement des fluctuations raisonnables (< 0,4 Å), contrairement aux systèmes monomériques où le pic ne diminue pas (*cf.* Figure 36). Cette diminution des RMSFs est facilement expliquée par le positionnement de cette région à l'interface des deux monomères constituant le dimère STAT5 (dSTAT5). En effet, ils interagissent de manière réciproque et forment un arrangement stable, limitant ainsi les fluctuations de ce segment.

Un pic important des RMSFs mais assez variable peut également être noté dans le DBD, au niveau des résidus 380 – 385. Ce pic correspond aux fluctuations des résidus situés sur la boucle reliant les brins *c* et *c'*. La visualisation de la structure de cette boucle très exposée au solvant a montré qu'elle peut parfois se détacher du corps du DBD avant de revenir se positionner contre la boucle reliant les brins *e* et *f*. Toutefois, et comme présenté par les courbes de RMSF, ce comportement n'est pas systématique, même au sein de la même simulation de dynamique moléculaire.

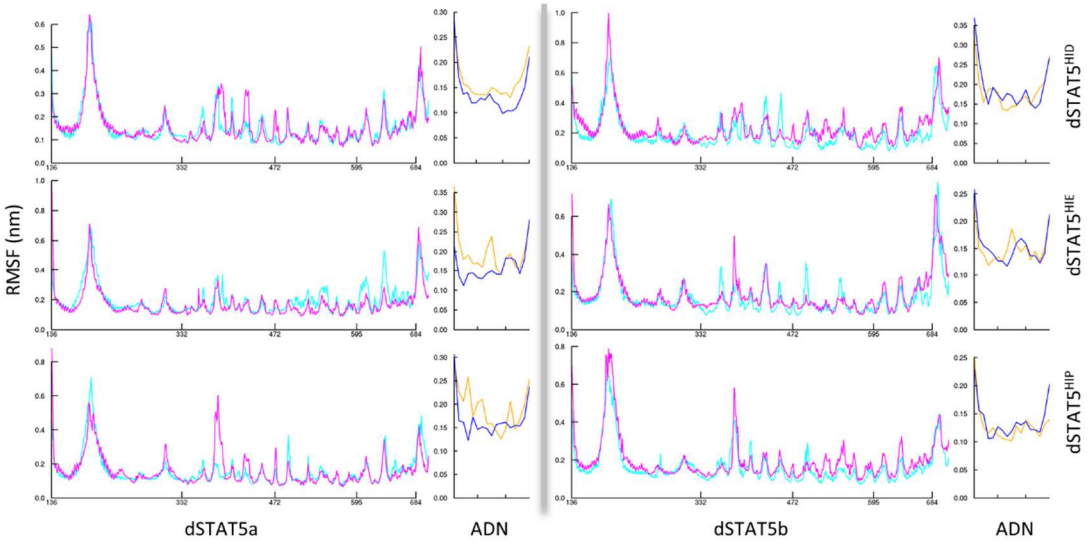


Figure 58: **Profils de déviation (RMSF) par résidu.** Les deux monomères de dSTAT5 sont représentés en cyan et en violet. Les deux brins d'ADN sont représentés en orange et en bleu. Les échelles sont différentes pour chaque STAT5/ ADN.

Les brins d'ADN présentent une dynamique très stable, au niveau central de la séquence de l'ADN, formée d'un double-brin. Cette stabilité s'exprime par des valeurs de RMSF inférieures à 2 Å, à l'exception des extrémités des brins qui présentent des valeurs plus élevées, mais jamais supérieures à 3,6 Å. Les variations de RMSD peuvent donc être expliquées par la flexibilité des extrémités des brins d'ADN, qui est facilitée par la forte exposition au solvant d'une part, par la

présence d'une base libre à l'extrémité 5' des brins d'autre part, et enfin par l'absence de contact avec la protéine donc de contraintes spatiales. La combinaison de ces trois conditions permet des fluctuations importantes des extrémités des brins comparativement aux résidus de la partie double-brin et au contact de STAT5.

B. Altération des structures secondaires

Comme pour les monomères, l'analyse des structures secondaires au cours des trajectoires des simulations de dynamique moléculaire a été réalisée afin de détecter les changements éventuels liés à la présence du double-brin d'ADN, des interfaces STAT5:STAT5 ou de la modification de l'état de protonation du résidu H471. L'arrangement des domaines de chaque monomère reste très similaire pour toute les simulations de dSTAT5 et même comparativement aux monomères. Cependant, des différences apparaissent (*cf.* Figure 59).

Au sein de chaque dimère de STAT5, des différences existent entre chaque monomère, indiquant un comportement dynamique qui n'est pas parfaitement identique, ce qui suggère l'absence de symétrique des mouvements des dimères. Ces observations corrèlent bien avec les courbes de RMSDs et RMSFs qui montrent également des différences entre les deux monomères de dSTAT5. Les plus grandes fluctuations sont trouvées dans la région de STAT5 qui assure le lien entre le domaine SH2 et la queue phosphotyrosyl, plus précisément entre la fin de l'hélice *αC* et le résidu de phosphotyrosyl. Ce segment de la protéine correspond à une boucle ne présentant peu structurée et dont les deux extrémités sont immobiles et agissent comme des points d'attache (*cf.* Figure 57). Les deux extrémités de ce segment sont d'une part l'hélice *αC*, en position N-terminale du segment, et d'autre part le résidu de phosphotyrosine, situé en position N-terminale. L'hélice *αC* est positionnée entre le feuillet β du domaine SH2 et l'hélice *αB* et présente une faible mobilité au cours des simulations de DM. Le résidu phosphotyrosyl est constamment lié au site d'interaction du monomère partenaire, et présente également une mobilité très faible. Entre ces deux éléments (hélice αC et résidu phosphotyrosyl), un segment de 30 ou 35 résidus pour STAT5a ou STAT5b, respectivement, porte de manière transitoire l'hélice αD ainsi que d'autres éléments variables (ponts ou brins β).

Ainsi, on peut voir entre les hélices *αC* et *αD* le prolongement du feuillet β du domaine SH2 et l'apparition d'un brin parallèle au brin *βA*, ou d'un pont β (résidus 669-670). Ces structures sont très variables et diffèrent d'un monomère à un autre. Par exemple, on observe le prolongement du feuillet β dans tous les monomères A de STAT5a, alors qu'un simple pont β est observé de manière transitoire dans les monomères B. Dans les dimères de STAT5b (dSTAT5b), un retrouve un repliement différent, caractérisé par le prolongement du feuillet β uniquement dans le monomère A de dSTAT5b^{HID}, alors que dans le monomère B, seul un pont β transitoire est retrouvé (*cf.* Figure 59). Dans les autres espèces de STAT5b, aucun prolongement du feuillet n'est observé, mais seulement des ponts β, qui peuvent être constants (monomère A de STAT5b^{HIE} et STAT5b^{HIP}) ou très transitoires (monomère B de STAT5b^{HIE} et STAT5b^{HIP}).

La structure de l'hélice *αD* varie au cours des simulations, pouvant adopter plusieurs arrangements en hélice de type α ou 3_{10} (cf. Figure 9). Le comportement de cette hélice semble similaire au sein de chaque dimère bien qu'aucun contact direct n'existe vu la distance qui les sépare (~ 40 Å). Ainsi, cette hélice présente une certaine dualité structurale – hélice α / hélice 3_{10} – dans les monomères des espèces dimériques de STAT5a. La courte hélice *αD* (résidus 675-678) présente un tour d'hélice variable au cours de la trajectoire de DM qui donne ainsi soit une hélice α soit une hélice 3_{10}. Le ratio hélice α / hélice 3_{10} est différent entre les monomères. Il est par exemple très déplacé vers l'hélice α (78 % du temps de simulation) dans le monomère B de dSTAT5a^{HIP} alors qu'il est équilibré dans le monomère A. L'hélice *αD* est suivie d'une hélice 3_{10} (résidus 679-682) relevée uniquement dans les 3 simulations de STAT5a. Chez STAT5b, le ratio hélice α / hélice 3_{10} diffère pour chaque monomère à l'exception de STAT5b^{HIP} qui montre le même ratio. Par comparaison avec dSTAT5a, l'hélice *αD* est détectée moins fréquemment, alors que nous n'observons pas la seconde hélice 3_{10} dans les dimères de STAT5b.

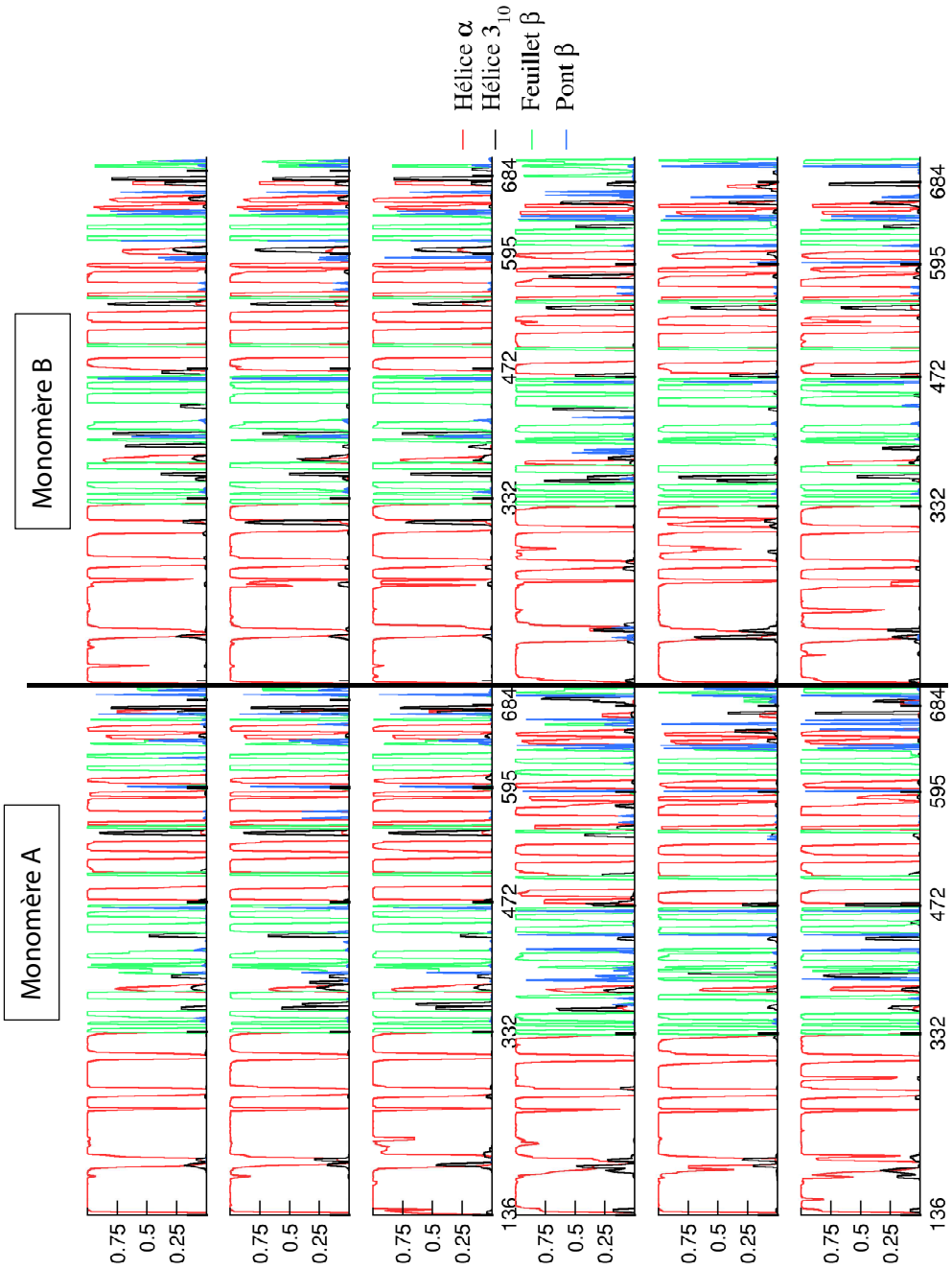


Figure 59: **Structures secondaires des dimères de STAT5.** Les deux monomères issus d'un même dimère sont représentés sur la même ligne. Les simulations correspondant à STAT5a^{HID}, STAT5a^{HIE}, STAT5a^{HIP}, STAT5b^{HID}, STAT5b^{HIE} et STAT5b^{HIP} sont présentées de bas en haut dans cet ordre. Les structures secondaires sont exprimées en probabilité de présence au cours des simulations.

Enfin, la partie C-terminale de la queue phosphotyrosyl montre des variations structurales importantes et est caractérisée par la formation d'un ou plusieurs feuillets β ou ponts β, voire d'une hélice 3_{10} (dans le monomère A de dSTAT5b^{HID}). Par exemple, dans la simulation de STAT5b^{HID}, le monomère A présente un pont β ainsi qu'un brin β, alors que le monomère B du même dimère montre deux brins β. Ces structures secondaires ne sont pas propres au monomère mais représentent des structures inter-monomères. Ainsi, le premier brin

β du monomère B est le prolongement du feuillet β du monomère A. Dans le cas du monomère A, un seul résidu forme des liaisons avec le feuillet, et est donc représenté par un pont β. Les derniers brins β de la séquence des deux monomères forment un feuillet β inter-monomère et interagissent l'un avec l'autre (*cf.* Figure 57). On peut voir dans le monomère B de STAT5b^{HID} que trois brins sont formés. Les deux premiers brins forment un feuillet antiparallèle dans une boucle de la queue phosphotyrosyl, et le dernier brin forme le feuillet inter-monomère (*cf.* Figure 60).

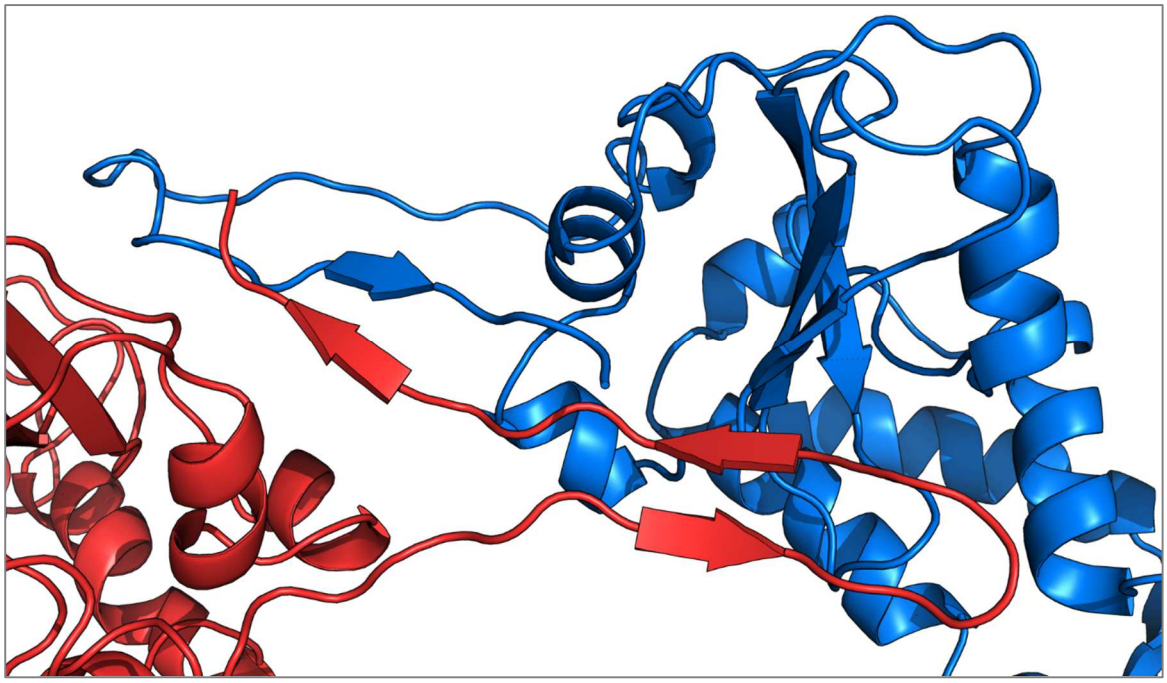


Figure 60: **Structure du domaine C-terminal du monomère B de STAT5b^{HID}.** Le monomère A est en bleu, le monomère B en rouge. Seuls les domaines SH2 et la queue phosphotyrosyl sont représentés pour plus de clarté.

Le tour de l'hélice αA est également raccourci dans les monomères B de dSTAT5a^{HIE} et dSTAT5a^{HIP} où une hélice 3_{10} est majoritairement détectée comparativement aux monomères A, alors que dans les simulations de dSTAT5b, elle est constamment détectée comme une hélice α.

D'autres variations dans la symétrie des structures secondaires sont à noter au sein du domaine DBD. Les monomères B des espèces dSTAT5a présentent ainsi une hélice 3_{10} dans la boucle reliant les brins *c'* et *e*, hélice qui n'est jamais retrouvée dans les monomères A. Dans les simulations de dSTAT5b, cependant, nous n'observons pas la même structure. Après contrôle des modèles de départ, les monomères B de dSTAT5a possèdent tous cette hélice 3_{10}, alors que les monomères A de dSTAT5a ou les monomères A et B de dSTAT5b ne la possèdent pas. Nous attribuons ainsi l'absence de symétrie entre les deux monomères de dSTAT5a à la différence de structure secondaire observée dans le modèle généré par homologie. Cependant, cette hélice est relativement stable puisque présente environ 70 % des temps de simulations. D'autres hélices 3_{10} ou α de ce domaine sont très variables d'une simulation à l'autre. Ainsi, une hélice 3_{10} couvrant les résidus 435 – 437 (soit quelques résidus avant le brin *f* du DBD) est observée plus de 50 % du temps de simulation dans le monomère A de dSTAT5a^{HIE} et dans le monomère B de

STAT5b^{HID}, mais est également retrouvée moins fréquemment dans les monomères A de dSTAT5a^{HID}, dSTAT5a^{HIP} et dSTAT5b^{HIP}. De même, d'autres hélices 3_{10} se forment sur la boucle entre les brins *b* et *c* (dSTAT5a^{HIE}, dSTAT5a^{HIP}, dSTAT5a^{HID}, monomère B de STA5b^{HIE} et dSTAT5^{HIP}), alors que l'hélice couvrant les résidus 377 à 379 et située juste après le brin *c* est absente du monomère B de STAT5b^{HIE} mais présente dans tous les autres monomères (*cf.* Figure 59).

Enfin, les hélices du domaine CCD montrent également des différences en termes de stabilité des structures secondaires. En particulier, les résidus 219-221, situés dans l'hélice *α2*, présentent une moindre occurrence (en % du temps de simulation) de stabilisation en hélice α pour le monomère B de dSTAT5b^{HIP} et dans une moindre mesure pour le monomère A de dSTAT5a^{HIP} (*cf.* Figure 59). Ces résidus correspondent à la région permettant l'articulation entre les deux modules dynamiques du CCD. Cette diminution d'occurrence de l'hélice *α2* à cette position particulière peut donc s'expliquer par la rupture transitoire de liaisons hydrogène au sein de l'hélice *α2* lorsque ces deux modules présentent des positions extrêmes. Les extrémités distales des hélices *α1* et *α2* voient dans certains systèmes la formation d'hélices 3_{10} en remplacement des hélices α, de manière plus marquée dans les dimères dSTAT5b (*cf.* Figure 59). Ces résultats corroborent les observations faites pour les espèces monomériques de STAT5 alors que nous n'observons pas de raccourcissement du brin *b* sur les formes dimériques. Le brin *C* du feuillet du domaine SH2 est moins bien formé dans les monomères issus des simulations de dSTAT5b que dans les systèmes dSTAT5a, mais présent alors que cela n'est pas le cas dans les formes de STAT5b monomérique.

Certaines spécificités liées à la séquence de STAT5 sont donc différentes dans les complexes dimériques liés à l'ADN (dSTAT5/ADN) et dans les formes de STAT5 monomérique. La présence d'interactions supplémentaires entre les deux monomères des dimères ou entre STAT5 et l'ADN peut être à l'origine de la modification structurale de ces protéines. Ainsi, le brin *C* du domaine SH2 se situe à l'interface entre les deux monomères avec qui des contacts stabilisants sont établis, à proximité des résidus de phosphotyrosine. Le feuillet β du domaine SH2 peut ainsi se prolonger en un brin *C* grâce au rapprochement des monomères.

C. Mouvements en ciseaux : une caractéristique de la famille des protéines STATs ?

La dynamique de toutes les formes dimériques de STAT5 semble similaire, comme suggéré par les profils de RMSD et RMSF. D'autre part, si des différences ont été notées en termes de structures secondaires, celles-ci révèlent davantage des changements à l'échelle locale (structures isolées) qu'à l'échelle du complexe complet. Afin de caractériser ces changements globaux, nous avons réalisé l'analyse en composante principale (ACP) des simulations de dynamique moléculaire. Les deux premières composantes principales de chaque simulation sont présentées dans la Figure 61.

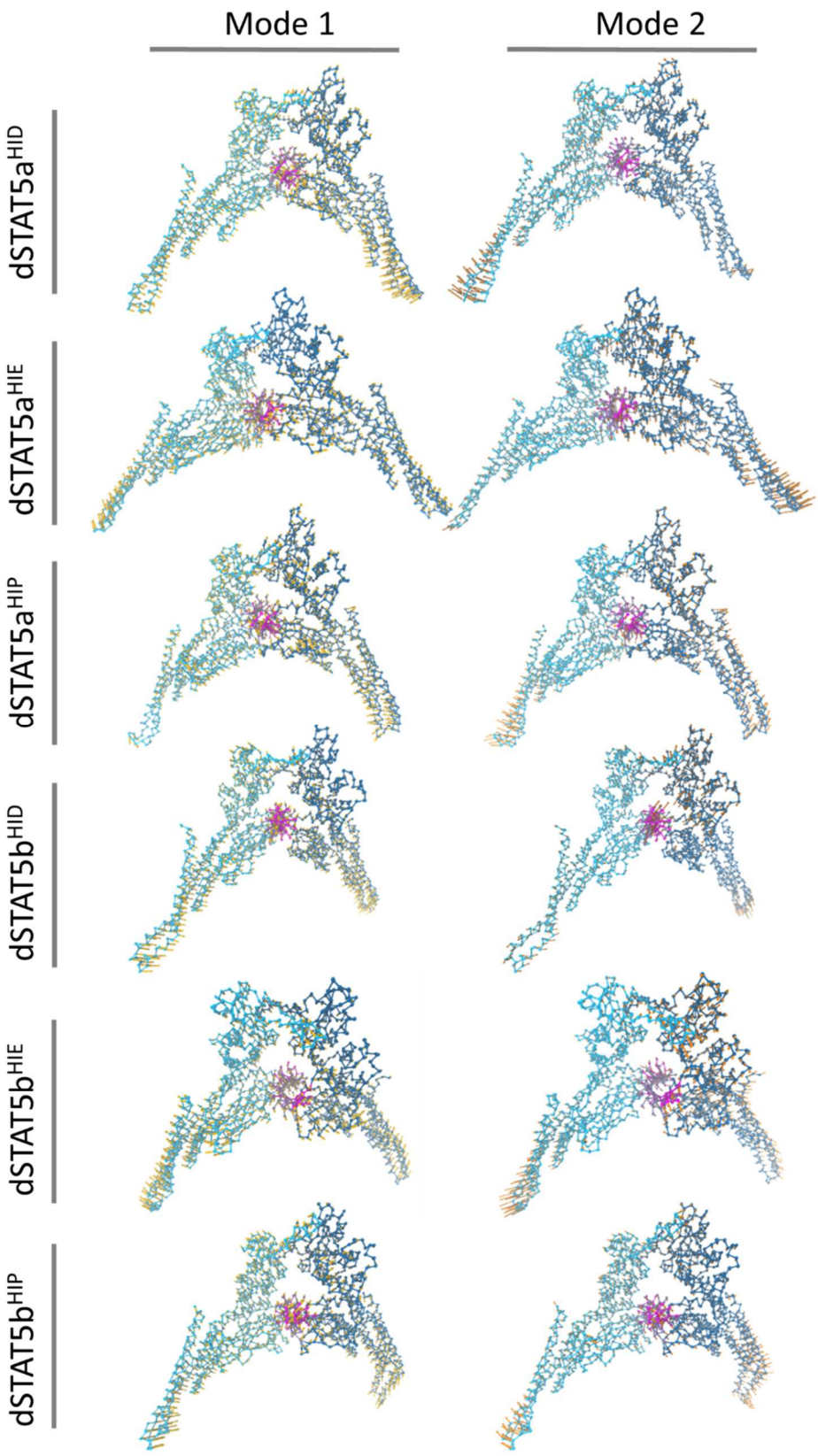


Figure 61: **Analyse des simulations de DM par ACP.** Les vecteurs associés aux deux premières composantes principales de chaque dimère sont affichés. Les monomères sont en bleu clair et bleu sombre, les brins d'ADN sont en magenta et violet.

Comme les profils de RMSF l'indiquent, les régions les plus fluctuantes des complexes dSTAT5/ADN sont les CCD distaux de chacun des monomères, ainsi que la partie de la queue phosphotyrosyl qui connecte le domaine SH2 au résidu phosphotyrosine. Le reste de cette région est quant à elle stable du fait de la présence d'interactions avec le domaine SH2 ou avec la queue phosphotyrosyl du monomère partenaire. Les vecteurs associés aux deux premières composantes principales (*cf.* Figure 61) décrivent des mouvements particulièrement larges et amples au niveau de la partie distale des CCDs de STAT5. Cependant, ces mouvements s'étendent à l'ensemble du domaine CCD, les amplitudes des déplacements augmentant en fonction de l'éloignement du centre géométrique du complexe, situé au niveau de la double-hélice d'ADN. La dynamique des dimères par rapport aux systèmes monomériques est donc plus sophistiquée : si on retrouve l'organisation en deux modules du CCD, ces mouvements sont associés à des mouvements globaux et concertés qui impliquent l'ensemble du domaine. On peut ainsi observer un gradient dans l'amplitude des mouvements, de la plus faible pour la partie des hélices α2-4 proche de l'ADN et du DBD à la plus grande pour la partie distale des hélices α1 et α2. La présence des mêmes types de mouvement dans toutes les dynamiques des deux isoformes de STAT5 semble indiquer que ce mouvement est davantage lié à l'architecture globale du complexe dSTAT5/ADN qu'à la séquence. Par ailleurs, le même type de mouvement a été décrit par Husby et collaborateurs dans leur étude du complexe STAT3/ADN[538], ce qui tend à montrer que cette dynamique est partagée par les protéines de la famille des STATs. Ce type de mouvement a été décrit comme similaire à ceux d'une paire de ciseaux (*scissor-like*), les deux lames représentant les domaines CCD, qui pivotent autour du double-brin d'ADN. Cependant, la symétrie de ces mouvements n'est pas parfaite et certains complexes étudiés présentent des différences notables entre les monomères (dSTAT5a^{HIP} par exemple, *cf.* Figure 61).

Les autres domaines (DBD, LD, SH2 et queue phosphotyrosyl) montrent des mouvements plus réduits par rapport de domaine CCD et différents d'un système à un autre. L'amplitude de ces mouvements dans ces domaines est similaire à celle du double-brin d'ADN. Les nucléotides présentent néanmoins une dynamique très variable en fonction de la région considérée, et ont logiquement tendance à adopter la dynamique du domaine protéique le plus proche (effet coopératif). Ainsi, les résidus nucléotidiques des grands sillons de l'ADN présentent une dynamique proche de la région du DBD qui s'insère dans le sillon, à savoir la boucle reliant le brin *g'* à l'hélice α5. Les extrémités des brins d'ADN montrent de plus grands déplacements, en accord avec les profils de RMSF et la présence de bases libres aux extrémités 5' de chaque brin.

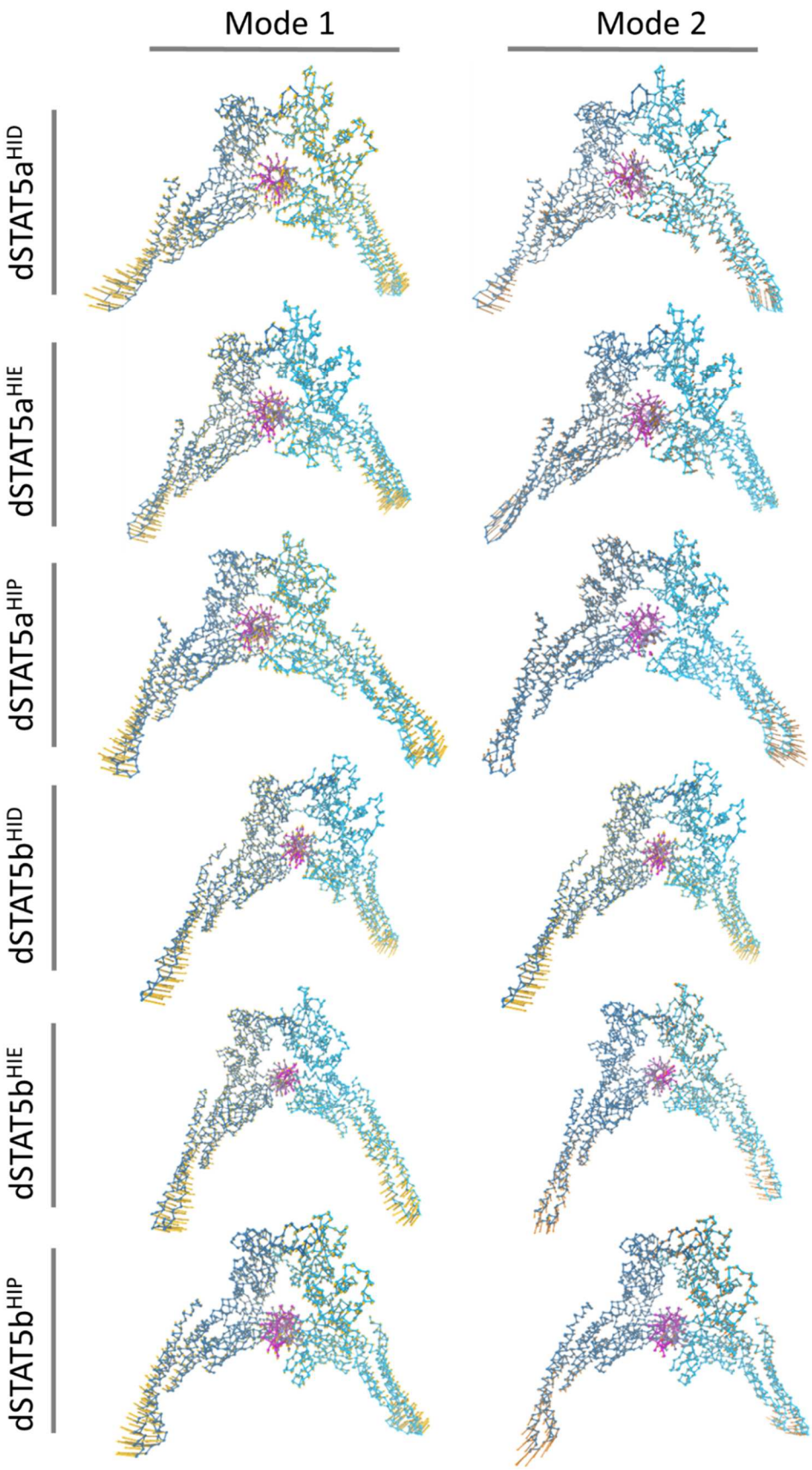


Figure 62: **Modes normaux des systèmes dimériques.** Pour chaque dimère, les deux premiers modes normaux sont affichés. Les monomères sont en bleu clair et bleu sombre, les brins d'ADN sont en magenta et violet.

Les premiers modes normaux présentent des caractéristiques très similaires pour tous les complexes étudiés, et les principaux traits des deux premiers modes sont semblables à ceux des premières composantes principales. On peut ainsi voir que le CCD montre les mêmes mouvements en ciseaux que nous avons observés au cours de dynamiques moléculaires, dont l'amplitude est maximum à l'extrémité distale. L'amplitude de ces mouvements est diminuée lorsque que les résidus sont proches de l'ADN.

Les autres domaines de STAT5 montrent des mouvements de plus faible amplitude. Les premiers vecteurs propres associés aux deux brins d'ADN montrent un mouvement très semblable à leur environnement protéique. Ainsi, les mouvements en ciseaux du CCD s'accompagnent du changement de la position des autres domaines : ils s'affaissent lorsque les CCDs s'écartent et inversement. Les domaines DBD, LD, SH2 et la queue phosphotyrosyl présentent ainsi une dynamique globale très homogène.

D. Interface protéine – ADN et influence de l'état de protonation de l'histidine 471

Le rôle et les fonctions physiopathologiques de STAT5 sont étroitement liés à sa capacité de reconnaître une séquence spécifique d'ADN, et à se lier à cette hélice à double-brin. Dans ce cadre, nous avons porté une attention particulière à l'interface protéine – ADN au sein des différents complexes dSTAT5 et nous avons recherché les contacts stabilisant le complexe et discriminer les interactions qui permettent d'expliquer la spécificité de reconnaissance ainsi que la liaison STAT5 – ADN.

La séquence d'ADN que nous avons modélisée est composée d'un double-brin d'ADN. Chaque brin est composé de 18 bases, dont une est libre à l'extrémité 5' alors que les autres sont engagées dans la formation d'un double brin de 17 paires de bases. Il a été montré que le motif le plus affin pour STAT5 est composé d'une structure palindromique de 9 bases dont la séquence est la suivant : $TTCN_3GAA$, où N_3 représente une succession de 3 bases dont la nature n'influe pas sur la spécificité de la reconnaissance STAT5 – ADN[530]. Nous avons donc modélisé ce motif et simulé les complexes afin de détecter les liaisons hydrogènes entre ces molécules. Les liaisons hydrogènes peuvent être considérées comme le partage d'un atome d'hydrogène entre deux atomes très électronégatifs (oxygène, azote ou soufre dans le cas des protéines). Nous avons utilisé des critères purement géométriques afin de détecter ce type de liaison au sein des complexes protéine – ADN (*cf.* paragraphe II.E.2 du chapitre 2).

Pour des raisons de clarté de la présentation et pour favoriser la comparaison de nos résultats aux données issues de la littérature, nous avons adopté la numérotation usuelle des résidus nucléotidiques qui place le nucléotide 0 au centre de la séquence palindromique. Ainsi, la séquence $TTCN_3GAA$ est numérotée de -4 à +4, et ainsi de suite pour les autres nucléotides. La numérotation des deux brins d'ADN est explicitée dans la Figure 63.

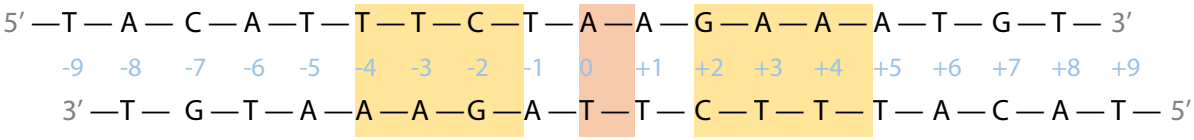


Figure 63 : **La séquence d'ADN utilisée pour modéliser les complexes STAT5/ADN et sa numérotation.** Les extrémités 3′ et 5′ sont indiquées, et les sites de reconnaissance spécifique sont sur fond jaune, alors que la paire de base centrale est sur fond orange.

L'analyse des structures des complexes dSTAT5/ADN permet établir de nombreux contacts non spécifiques, entre les résidus protéiques et la chaîne principale de l'ADN ou les oses (*cf.* Figure 63) et non la partie base azotée du nucléotide. De nombreuses liaisons sont détectées mais plusieurs d'entre elles ne sont retrouvées que dans quelques monomères de dSTAT5, indiquant une large variabilité des liaisons hydrogènes à la surface de l'ADN. Trois résidus (K343, R423 et K425) sont impliqués dans ce type d'interaction dans tous les systèmes, dont deux (K343 et R423) dans tous les monomères, le résidu K425 du monomère B de dSTAT5a^{HIE} ne formant pas de liaison hydrogène avec l'ADN. En réalité, ce résidu forme bien des liaisons hydrogène avec la chaîne principale de l'ADN, mais pendant des durées inférieures à 6% du temps de simulation, ce qui l'exclut de la Figure 63. Ainsi, malgré la présence d'une forte variabilité des liaisons hydrogènes détectées, les résultats indiquent que les résidus K343, R423 et K425 sont les principaux acteurs de la liaison aux chaînes principales de l'ADN. La position de ces résidus, sur une boucle flexible et leur nature similaire (ces trois résidus sont chargés positivement), expliquent la forte occurrence de leurs liaisons avec les atomes de la chaîne principale étant chargée négativement.

Tous les résidus nucléotidiques sont contactés par dSTAT5, à l'exception des deux ou trois résidus situés en 5′ des brins, ainsi que du motif AA situé position +4 et +5, ou -4 et -5 en fonction du brin considéré. La symétrie des contacts n'est pas parfaite, même si elle est souvent retrouvée. Certains contacts dSTAT5 –ADN observés au cours des dynamiques moléculaires ne sont pas relevés dans la structure cristallographique 1BG1[132], que nous avons utilisé comme support pour la modélisation par homologie. Ces contacts retrouvés à partir des données de simulation DM concernent surtout les résidus nucléotidiques positionnés aux extrémités des double-brins (aux positions ±6, ±7 ou ±8, *cf.* Figure 63), mais également le résidu R423 correspondant au résidu E 416 chez STAT3, notamment. Ces résultats sont en accord avec l'étude de dynamique moléculaire menée sur le complexe STAT3 / ADN[538] qui note aussi une asymétrie des interactions STAT – ADN.

***Tableau 11* : Résidus protéiques impliqués dans la formation de liaisons hydrogènes protéine - ADN non spécifiques.** Les résidus de protéine contactant la chaîne principale ou les oses des résidus nucléotidiques observés pendant plus de 10% du temps de simulation sont répertoriés ci-dessous. Les liaisons hydrogènes présentes pendant plus de 50% du temps sont indiquées en gras. Une liaison hydrogène observée entre le résidu X des monomères A et B sera indiquée par A/B, alors que si la liaison n'est observée qu'entre un seul monomère, elle sera marquée par A ou B.

		dSTAT5a			dSTAT5b			
		HID	HIE	HIP	HID	HIE	HIP	ADN
Boucle *a' b*	K343	**A**/B	A/B	**A**/B	**A/B**	**A**/B	A/**B**	±1, 0, +2
	T344		B	B				-5, +4
	Q345				A			-3
	T346	A/B				A	**A/B**	0
	K347	**B**		**B**	**A/B**	A/B	**A/B**	-2, ±1, 0
Boucle *c c'*	N390					**A**		-5
	S393						B	+4
	Q395		B					0
Boucle *e f*	R417				A/B	A	B	±6
	N418				A/B			±2
	K422		B		B			-1, +3
	R423	A/B	**A/B**	**A/B**	**A/B**	**A/B**	**A/B**	+3, ±4, ±5
	K425	**A**/B	**A**	A/B	A/B	A/B	**A**/B	±8, +6, +7, ±3, ±4
	R426	**A**/B	A/B	A			**A**	+2, ±3, ±4
	A/S427		**B**			B	**A**	±4, +7
	D428					A		-8
	R429	A		B	A		A/**B**	±8
	R430	B		A		B	**A/B**	-5, +6, ±8
	G431	B						+6
	E433					B	B	+8
	S434	B	**B**	A/**B**		B		+5, ±6
	V435	**B**		A/**B**		B	**A/B**	±5
	T436		**A/B**	**B**		B	**B**	±5, +6
Boucle *g' α5* et hélice *α5*	H471		B	**A**	A/B			-4, ±5, +6
	G472						A/B	±1
	S473	B	A	A		A		-1, 0, ±2
	Q474		B		A/B	**B**	B	±5, +7
	N477					B		+5

		dSTAT5a			dSTAT5b			
		HID	HIE	HIP	HID	HIE	HIP	ADN
Boucle α7''α8	R560		**A**	A				-6
	N567					B		+7
Hélice α8	K582					B		+1
	K583	**A**	A	A	A/B	B		0, ±1, +2
Hélice αB	R649	A/B	A	A/B	A/**B**	**B**		-1, ±2, +3
	S652			A				+3

Ces données mettent en avant capacité de STAT5 à s'adapter à l'ensemble de la surface de l'ADN et à former divers contacts. Cependant, ces liaisons hydrogènes n'expliquent pas la spécificité de séquence de STAT5 pour la reconnaissance de ce motif d'ADN car elles n'impliquent aucun atome des bases azotées. Nous avons réalisé la même analyse en nous focalisant sur les atomes des bases azotées afin de détecter les liaisons hydrogènes responsables de la reconnaissance STAT5 – ADN (*cf.* Tableau 12).

Tableau 12: **Liaisons hydrogènes entre les résidus protéiques et les bases azotées.** Les liaisons hydrogènes qui impliquent les bases du motif $TTCN_3GAA$ sont indiquées par le fond bleu. Les atomes NH1 et NH2 correspondent aux atomes d'azote situés à l'extrémité de la chaîne latérale des résidus d'arginine, les atomes O correspondent à l'atome d'oxygène de la chaîne principale des résidus, les atomes Nδ et Nε correspondent aux atomes d'azote positionnés en δ et en ε dans les résidus d'histidine.

Résidu - atome (monomère)	Base azotée (position) - atome	Occurrence (en % du temps de simulation)	Systèmes dimériques
R429 – NH1 (B)	A(+6) – N3	25,3	$dSTAT5a^{HID}$
H471 – Nε (B)	A(-2) – N6	22,7	$dSTAT5a^{HID}$
R429 – NH1 (B)	A(+6) – N3	86,7	$dSTAT5a^{HIE}$
R429 – NH2 (B)	A(+5) – O2	61,8	$dSTAT5a^{HIE}$
D428 – O (A)	T(-8) – N3	12,6	$dSTAT5a^{HIE}$
H471 – Nε (A)	G(-2) – O6	28,8	$dSTAT5a^{HIE}$
H471 – Nε (B)	T(-4) – O4	10,0	$dSTAT5a^{HIE}$
R429 – NH1 (B)	A(+6) – N3	40,0	$dSTAT5a^{HIP}$
R429 – NH2 (B)	T(+5) – O2	27,8	$dSTAT5a^{HIP}$
K425 – O (B)	T(+8) – N3	10,3	$dSTAT5a^{HIP}$
H471 – Nε (B)	G(-3) – N7	37,3	$dSTAT5a^{HIP}$
H471 – Nε (B)	G(-3) – 06	25,8	$dSTAT5a^{HIP}$
R429 – NH1 (A)	A(-8) – N3	46,4	$dSTAT5b^{HIE}$
R429 – NH1 (A)	C(-7) – O2	12,7	$dSTAT5b^{HIE}$
R429 – NH2 (A)	C(-7) – O2	78,4	$dSTAT5b^{HIE}$
H471 – Nε (B)	G(+2) – O6	64,8	$dSTAT5b^{HIE}$
R429 – NH1 (A)	T(-8) – O2	14,7	$dSTAT5b^{HIP}$
H471 – Nδ (A)	G(-2) – O6	10,3	$dSTAT5b^{HIP}$

Résidu - atome (monomère)	Base azotée (position) - atome	Occurrence (en % du temps de simulation)	Systèmes dimériques
H471 – Nε (A)	T(-4) – O4	28,6	dSTAT5b^{HIP}
H471 – Nε (B)	G(+2) – O6	43,5	dSTAT5b^{HIP}
H471 – Nε (B)	T(+4) – O4	20,0	dSTAT5b^{HIP}
H471 – Nε (B)	T(+3) – O4	18,6	dSTAT5b^{HIP}

L'analyse détaillée des contacts livre plusieurs éléments intéressants. Tout d'abord, et bien que la reconnaissance STAT5–ADN ait été isolée sur un motif bien déterminée, les liaisons hydrogènes spécifiques (nous emploierons le terme de spécifique pour désigner les liaisons hydrogènes entre un résidu protéique et un atome d'une base azotée) ne concernent pas que les résidus de ce motif, aux positions ±2, ±3 et ±4 (Tableau 12**)**. Des liaisons sont ainsi observées avec les bases des nucléotides situées aux positions +5, +6, -7 et ±8. Ehret et collaborateurs ont bien noté des variations de l'affinité de STAT5 pour l'ADN en lien avec la nature de ces bases, mais elles restent néanmoins peu marquées et ne peuvent être qualifiées de spécifiques[530]. Enfin, aucun contact de STAT5 avec l'ADN n'est associé aux nucléotides situés au centre du motif de reconnaissance, situés aux positions ±1 et 0. Ces données corrèlent bien avec l'absence de spécificité de reconnaissance au niveau de ces résidus.

Les liaisons hydrogènes observées sont présentes de manière principalement transitoire : la majorité n'est détectée que moins de 50% du temps de simulation. Quatre liaisons sont observées au-delà de cette limite, et seule une liaison hydrogène implique une base du site de reconnaissance de l'ADN par STAT5. Si la faible occurrence des interactions spécifiques avec les bases en dehors du motif de reconnaissance semble corréler aux données expérimentales, la faible occurrence (voire l'absence) de liaisons hydrogènes spécifiques (dSTAT5b^{HID}) avec les nucléotides porteurs de la spécificité semble contradictoire avec la fonction de reconnaissance spécifique l'ADN par STAT5.

Plusieurs facteurs peuvent apporter un élément de réponse. Tout d'abord, les liaisons hydrogènes ne sont pas les seules interactions qui apportent la spécificité dSTAT5 / ADN, la reconnaissance d'un motif d'ADN par une protéine résultant d'un ensemble d'interactions de différents types[617]. La reconnaissance d'une séquence d'ADN est ainsi la combinaison de la reconnaissance de la forme du double-brin d'ADN (courbure, ADN A, *etc.*) la reconnaissance des bases azotées. La « lecture » des bases azotées d'un grand sillon est principalement réalisée apr. la formations de liaisons hydrogènes, mais peut également se faire par la formation de contacts hydrophobes ou de contacts *via* les molécules d'eau[617]. Les liaisons hydrogènes constituent néanmoins le principl vecteur de la reconnaissance des bases azotées de l'ADN par les protéines. Ensuite, le dimère dSTAT/ADN a été modélisée à partir d'une structure de STAT5 murin (code PDB : 1Y1U) et d'un homodimère de STAT3 (code PDB : 1BG1). Dans la première structure, l'absence d'ADN permet au résidu d'histidine en position 471 d'adopter une orientation favorable énergétiquement qui ne correspond pas nécessairement à celle qu'elle adopterait en présence d'ADN. L'analyse de cette structure révèle ainsi que la chaîne latérale de

l'histidine 471 est tournée en direction d'une hélice positionnée juste avant le brin *f* du domaine de liaison à l'ADN. Cette position est à l'opposé de la position potentielle qu'elle adopterait si, comme les données biologiques le suggère, elle établit des contacts spécifiques avec l'ADN. La qualité des modèles peut donc être biaisée sur ce point. De même, le résidu correspondant à l'H471 dans la structure 1BG1 est un résidu d'asparagine (N466). La différence de résidu peut là aussi avoir introduit un biais qualitatif au cours de l'étape de modélisation par homologie. L'étude de dynamique moléculaire réalisée à partir de cette structure a montré que ce résidu N466 de STAT3 établit des liaisons hydrogènes spécifiques stables (>90% d'occurrence au cours d'une dynamique du même ordre de longueur que nos simulations) avec les bases azotées aux positions ±2 et ±3[538]. Ce résultat semble indiquer que les interactions spécifiques s'établissent rapidement au cours de la dynamique et que le temps d'équilibration ne constitue pas un facteur limitant. Cependant, il pourrait être utile de prolonger certaines simulations afin de s'assurer de cet effet. La nature du résidu est également un élément qui peut expliquer la baisse d'occurrence des liaisons hydrogènes spécifiques comparativement au complexe STAT3 : ADN. Le résidu N466 possède sur sa chaîne latérale un atome d'oxygène, accepteur de liaison hydrogène, mais cet atome ne semble impliqué dans la formation de liaisons hydrogènes. L'atome d'azote de la chaîne latérale du résiduN466 peut former simultanément deux liaisons hydrogène[538], là où les résidus d'histidine ne peuvent partager qu'un seul proton. Ainsi, et même dans les formes HIP des dimères de STAT5 où la position des deux atomes d'azote est incompatible avec la présence de deux liaisons d'hydrogène, il n'est pas étonnant de noter une différence significative en termes de stabilité des interactions résidu / ADN entre les résidus H471 de STAT5 et N466 de STAT3. La différence de comportement entre les différentes protonations de H471 peut s'expliquer par la position des atomes d'azote sur le cycle imidazole. L'atome d'azote protoné en position ε est ainsi dirigé vers les bases azotées dans nos modèles, ce qui favorise la formation des liaisons hydrogènes spécifiques. *A contrario*, l'atome d'azote δ est plus difficilement en contact les bases azotées du fait de l'encombrement du cycle imidazole.

Ces résultats indiquent des différences notables de la capacité de l'histidine H471 à former des liaisons spécifiques avec l'ADN dans différents états de protonation. Si nous avons observé des contacts spécifiques dSTAT5/ADN avec l'ensemble des résidus porteurs de la spécificité de liaison, ces interactions ne sont en général pas stables (<45% du temps de simulation) à l'exception d'une liaison survenant environ 65% du temps de simulation dans le système dSTAT5b^{HIE}. Lorsque toutes les liaisons sont prises en compte, aucune symétrie dans les interactions n'est observée, quel que soit le dimère de STAT5 considéré, à l'image des interactions non-spécifiques. Enfin, des interactions spécifiques entre STAT5 et la partie de l'ADN qui ne porte pas la spécificité de reconnaissance sont également observées. Le rôle de ces interactions dans la formation des complexes STAT5 : ADN reste à déterminer.

E. Interface STAT5–STAT5 et positionnement de la queue phosphotyrosyl

La phosphorylation du résidu de tyrosine, Y694 (STAT5a) ou Y699 (STAT5b), est un évènement crucial dans le cycle d'activation de STAT5 car il autorise le passage de la forme monomérique (cytoplasmique libre) à la forme dimérique parallèle capable de se lier à l'ADN. La liaison d'un groupement phosphate $-PO3^{2-}$ sur la chaîne latérale d'un résidu tyrosine permet d'augmenter considérablement la formation de multiples liaisons polaires, notamment des liaisons hydrogènes favorisées et stabilisées par le phosphotyrosine. La présence de ce groupe chimique apporte trois atomes d'oxygène dont l'électronégativité est supérieure à celle du groupe hydroxyl du résidu de tyrosine non phosphorylé. L'ajout de ce groupe chimique explique donc le gain de fonction observé pour les formes phosphorylées des protéines STAT.

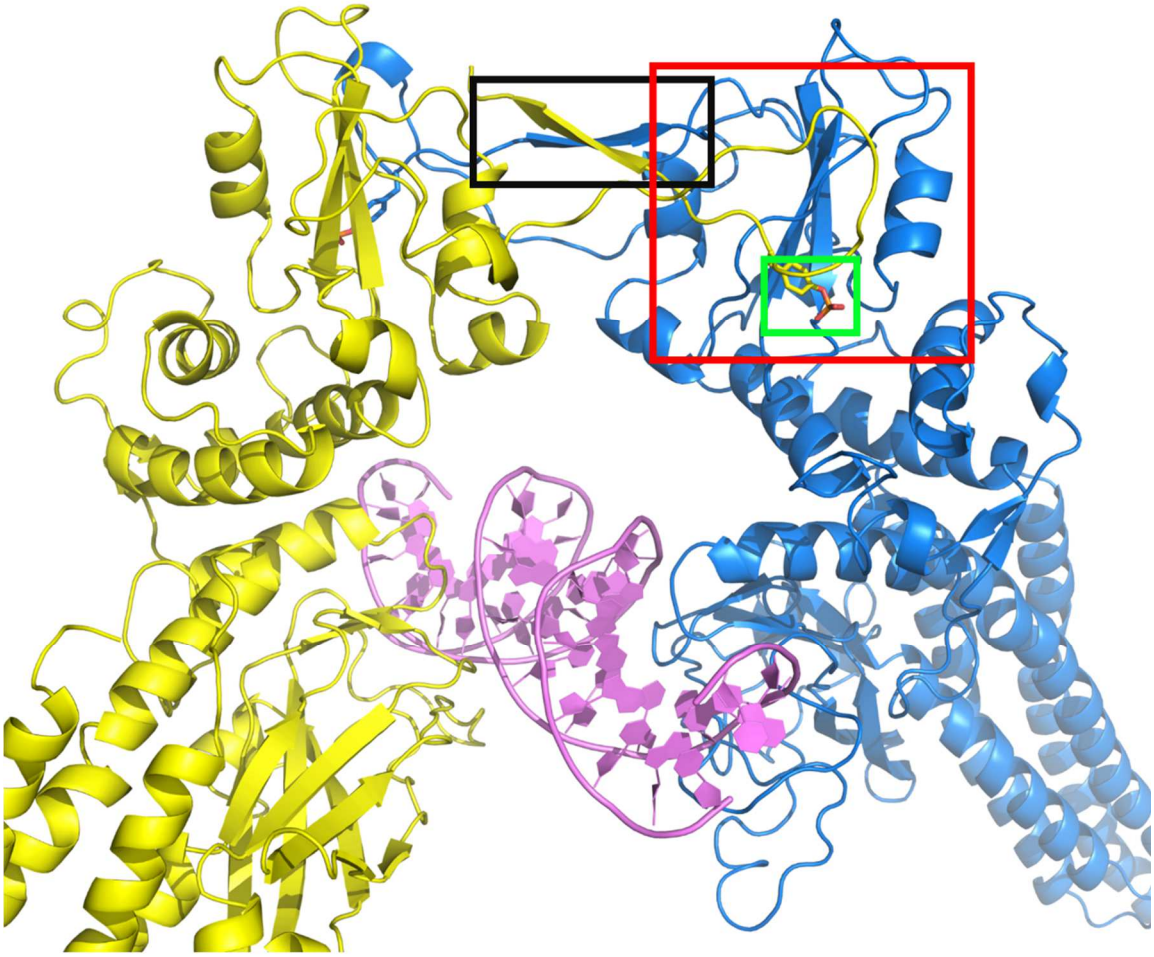


Figure 64 : **Vue générale de l'interface entre les monomères du complexe STAT5/ADN.** Les deux monomères de STAT5 sont représentés en jaune et en bleu, l'ADN en violet. Les trois principaux sites d'interactions monomère:monomère sont encadrés en vert (site de liaison de la phosphotyrosine), en noir (formation de structures secondaires inter-monomères) et en rouge. Le dimère représenté est dSTAT5b^{HIP}.

Si l'interface STAT5 : STAT5 est principalement caractérisée par les interactions liées au résidu de phosphotyrosine, d'autres interactions sont également présentes de manière stable comme l'ont révélé les résultats précédents (présence d'un feuillet β à l'interface monomère /

monomère, *cf.* paragraphe II.B de ce chapitre). Afin de caractériser plus finement ces interactions nous avons réalisé l'analyse des liaisons hydrogènes entre les deux monomères qui constituent les complexes dSTAT5 / ADN. Trois principaux sites sont impliqués dans les interactions monomère:monomère (*cf.* Figure 64).

Nous avons observé que le résidu de phosphotyrosine, du fait de la présence du groupement phosphate, est capable de former un réseau très dense de liaisons hydrogènes (*cf.* Tableau 13 et Figure 65). Cinq de ces liaisons, impliquant K600, R618, S620, D621 et S622, sont présentes de manière quasi-constante dans tous les systèmes simulés. Si l'implication de K600, R618, S620 et S622 était attendu au vu des données cristallographiques disponible pour les dimères de STAT3 et STAT1 dans lesquelles les résidus correspondant sont répertoriés comme susceptibles de former ces liaisons, le rôle de D621 (ou des résidus correspondant E605 ou E612 dans STAT1 et STAT3, respectivement) dans la liaison entre la phosphotyrosine et le domaine SH2 n'a à notre connaissance jamais été évoqué. Le résidu D621 est le seul résidu non-conservé au sein de la famille des STATs. La liaison hydrogène le liant au résidu de phosphotyrosine implique l'atome d'azote de la chaîne principale (*cf.* Figure 65). Les résidus correspondants à D621 au sein des complexes des autres protéines STATs pourraient donc également être impliqués dans les interactions entre la phosphotyrosine et le domaine SH2. Au contraire des liaisons hydrogènes observées entre l'ADN et STAT5, ces cinq liaisons hydrogènes formées par la phosphotyrosine présentent un caractère symétrique et sont présentes dans chaque paire de monomères. Ainsi, la dynamique du complexe ne semble que peu influer sur l'étroitesse des interactions entre ces sites.

Tableau 13: **Résidus formant des liaisons hydrogènes avec le groupement phosphate du résidu de phosphotyrosine.** Pour les deux monomères de chaque simulation, l'occurrence de la liaison hydrogène est indiquée en pourcentage du temps de simulation. Seules les liaisons présentes plus de 10% du temps de simulation sont indiquées.

	dSTAT5a			dSTAT5b		
	HID	HIE	HIP	HID	HIE	HIP
K600	93/100	95/93	95/93	85/57	73/81	19/73
R618	100/100	100/100	100/100	87/100	97/100	100/100
S620	67/99	99/83	97/97	59/96	30/85	99/95
D621	96/97	87/97	94/98	73/88	81/98	97/95
S622	98/100	100/99	99/94	98/85	92/100	100/99
T628	-/-	-/-	-/-	-/-	-/-	99/-

Enfin, le résidu T628, dans un monomère du système dSTAT5b^{HIP}, a montré une forte capacité à lier le groupement phosphate *via* sa fonction hydroxyle. Cette liaison n'est observée dans aucun autre système. Ce phénomène est lié à la rotation du groupement phosphate autour de la liaison entre le groupement phényl et le groupement phosphate de la chaîne latérale, ce qui a pour effet de diriger ce groupement vers le résidu de thréonine. Ce résidu n'a à notre connaissance jamais été présenté dans la littérature comme crucial pour la liaison de la phosphotyrosine. La rotation du groupement phosphate serait donc un évènement fortuit, et la

liaison T628 / phosphotyrosine une conséquence de cet évènement non systématique pour les protéines dSTAT5. Néanmoins, ce résultat tend à souligner que d'autres résidus pourraient jouer un rôle dans la formation de la liaison entre les résidus de phosphotyrosine et les domaines SH2.

En dehors du site de liaison de la phosphotyrosine, d'autres interactions sont observées, notamment au niveau de l'extrémité C-terminale de nos modèles. De nombreuses liaisons hydrogènes sont observées dans les dimères dSTAT5, mais il ne semble pas y avoir d'impact de l'état de protonation du résidu H471 (*cf.* Tableau 14). Trois liaisons hydrogènes sont néanmoins observées presque constamment à l'interface des homo-dimères de STAT5b et impliquent les résidus Q703, K705 et V707 (*cf.* Figure 66). Dans le cas de STAT5a, seule la liaison centrale entre les deux résidus K700 de chaque paire de monomères (correspondant aux résidus K705 dans STAT5b) est fréquente alors que les autres liaisons hydrogènes ne sont pas observées dans tous les dimères de STAT5. D'autres liaisons hydrogènes sont observées entre différents résidus de STAT5s, mais sont variables d'un complexe à un autre. La variabilité de ces liaisons permet d'expliquer en partie la fluctuation des structures secondaires observées au niveau de cette région et présentée dans le paragraphe II.B de ce chapitre.

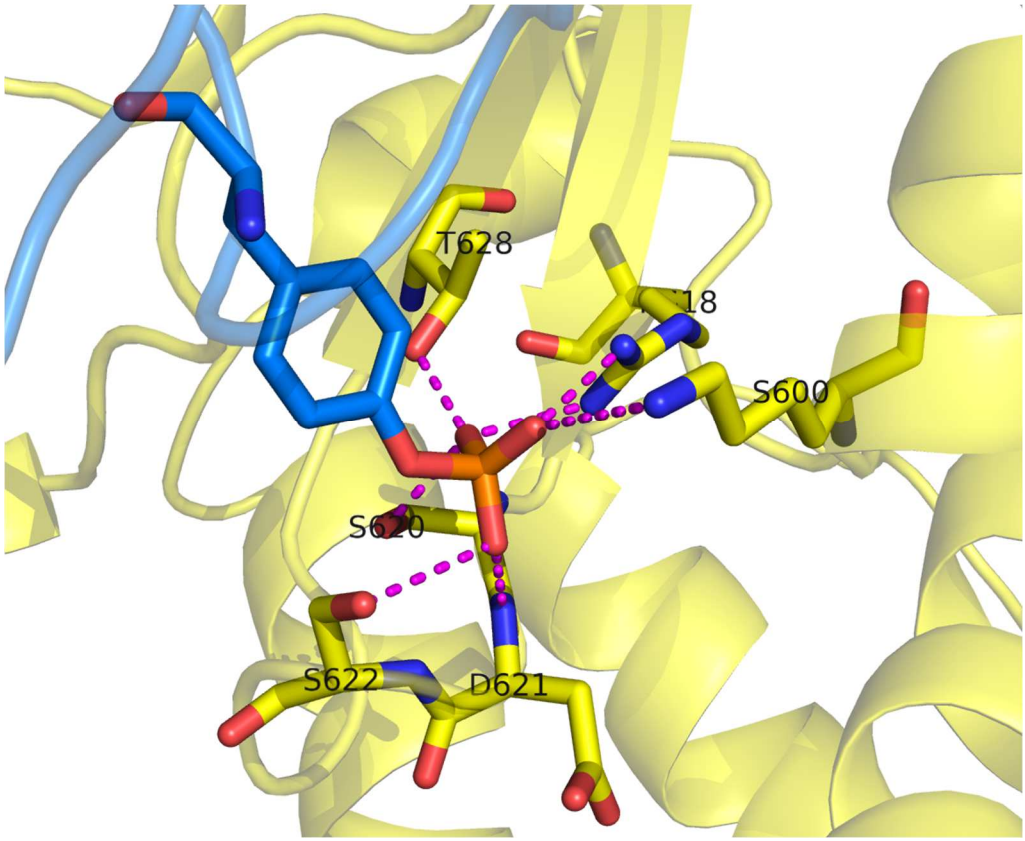


***Figure 65 :* Réseau de liaisons hydrogènes formé par le résidu de phosphotyrosine.** Les liaisons hydrogènes sont représentées en pointillés magenta. Un monomère est coloré en jaune, le second en bleu. Le système présenté est dSTAT5b^{HIP}.

Tableau 14: **Liaisons hydrogènes observées entre les deux extrémités de la queue phosphotyrosyl.** La double numérotation des résidus vient de l'insertion CESAT dans STAT5b. L'occurrence de chaque liaison est indiquée pour tous les monomères des simulations. Seules les liaisons présentes plus de 10% du temps de simulation sont indiquées.

Accepteur	Donneur	dSTAT5a			dSTAT5b		
		HID	HIE	HIP	HID	HIE	HIP
P697/702 (O)	Q701/706 (NE2)	-/30	-/48	-/60	-/-	-/-	-/-
Q698/703 (O)	Q701/706 (NE2)	-/13	-/-	-/-	-/-	-/-	-/-
Q698/703 (O)	V702/707 (N)	-/53	-/60	-/-	98/56	99/59	93/97
K700/705 (O)	K700/705 (N)	97/38	96/58	95/-	58/93	30/98	95/92
Q701/706 (O)	Q698/703 (N)	-/95	-/81	16/-	-/-	-/-	-/-
Q701/706 (O)	I699/704 (N)	-/-	-/18	-/-	-/-	-/-	-/-
Q701/706 (O)	K700/705 (NZ)	16/-	-/-	60/-	-/20	-/-	-/-
Q701/706 (OE1)	K701/706 (N)	-/-	-/-	-/-	39/-	-/-	-/-
V702/707 (O)	Q698/703 (N)	54/-	33/15	48/-	56/97	98/98	92/89
V703/708 (OC1)	K700/705 (NZ)	-/-	-/-	27/-	-/-	-/-	-/-
V703/708 (OC2)	K700/705 (NZ)	-/-	-/-	15/-	-/-	-/-	-/-

L'analyse du réseau de liaisons hydrogènes peut laisser penser à la présence d'un effet dépendant de la séquence au vu de l'existence constante de deux liaisons hydrogènes chez STAT5b, qui ne sont présentes de manière peu fréquente dans dSTAT5a. Pour confirmer cette observation, une analyse plus approfondie de cette interface est nécessaire afin de définir plus précisément si les différences du réseau de liaison hydrogène que nous observons se reflètent dans l'énergie de liaison entre les deux monomères.

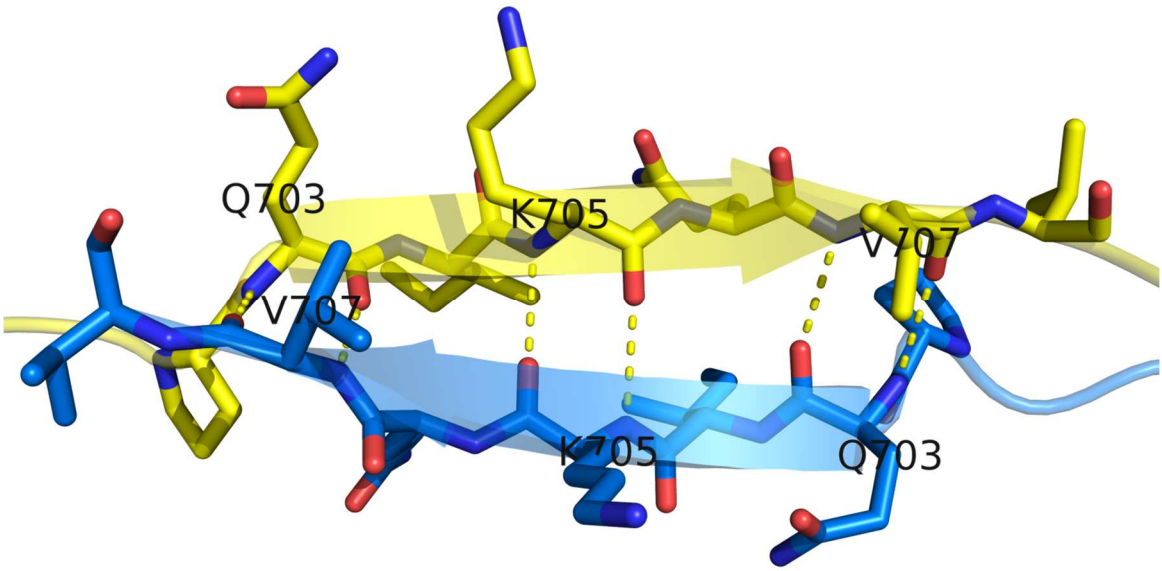


Figure 66 : **Liaisons hydrogènes entre les extrémités C-terminales des protéines dSTAT5.** Les liaisons hydrogènes sont représentées en pointillés jaune. Un monomère est coloré en jaune, le second en bleu. Le système présenté est dSTAT5b^{HIP}.

D'autres liaisons hydrogènes sont détectées à proximité du site de liaison de la phosphotyrosine, ou à l'interface entre la queue phosphotyrosyl et le domaine SH2. Précisément, le résidu valine, V695 (STAT5a) ou V700 (STAT5b), situé après la phosphotyrosine forme des liaisons hydrogènes avec les résidus N642 et K644 (STAT5a) ou M644 (STAT5b) du brin *C* du domaine SH2 du second monomère (*cf.* Tableau 15 et Figure 67). Ces liaisons additionnelles à celles de la phosphotyrosine viennent ainsi stabiliser l'interface entre les monomères de STAT5 en offrant un second point d'ancrage, et permettent de maintenir la queue phosphotyrosyl et la phosphotyrosine dans une position stable à proximité de sont site de fixation complémentaire, le domaine SH2.

D'autres liaisons hydrogènes sont observées, de manière très variable et transitoire dans la plupart des cas. Néanmoins, elles impliquent un nombre limité de résidus en plus des résidus N642, K/M644 et V695/700, en fonction de la protéine et de l'état de protonation de l'histidine considérée (*cf.* Figure 67). Ainsi, dans la simulation de dSTAT5b^{HIP}, les résidus K600, Q636, W641, A691, A693, K694, V696, D697 et G698 sont impliqués à différents moments de la simulation dans des liaisons hydrogènes inter-monomères.

Tableau 15 : **Autres liaisons hydrogènes entre les monomères des dSTAT5.** Les deux types de résidus sont indiqués lorsque les résidus sont changés entre STAT5a et STAT5b. La double numérotation des résidus vient de l'insertion CESAT dans STAT5b. L'occurrence de chaque liaison est indiquée pour tous les monomères des simulations. Seules les liaisons présentes plus de 10% du temps de simulation sont indiquées.

Accepteur	Donneur	dSTAT5a			dSTAT5b		
		HID	HIE	HIP	HID	HIE	HIP
E623 (OE1)	K/M644 (NZ)	-/-	-/-	-/12	-/-	-/-	-/-
E623 (OE2)	K/M644 (NZ)	-/-	-/-	-/19	-/-	-/-	-/-
N/M639 (ND2)	V691/696 (O)	10 /-	-/-	-/-	-/-	-/-	-/-
N642 (N)	G693/698 (O)	67/-	-/-	-/-	-/-	-/-	-/-
N642 (O)	V695/700 (N)	94/97	93/99	96/97	24/96	96/75	96/95
D650 (OD1)	K701 (NZ)	-/-	-/-	-/-	17/-	-/-	-/-
I653 (O)	R659 (NE)	11/-	-/-	-/-	-/-	-/-	-/-
A688/693 (O)	Q636 (NE2)	-/-	-/-	-/-	-/-	-/-	12/-
A686/691 (O)	W641 (NE1)	-/-	-/-	-/-	-/-	-/-	-/70
A688/693 (O)	W641 (NE1)	-/-	-/-	-/-	-/-	-/-	19/-
K689/694 (O)	Q636 (NE2)	-/-	-/-	-/-	-/-	-/-	13/-
K689/694 (O)	K600 (NZ)	-/-	-/-	-/-	-/-	12/-	-/-
V691/696 (O)	K600 (NZ)	-/-	-/-	-/-	-/-	-/-	27/-
D692/697 (OD1)	R638 (NE)	-/-	-/-	-/-	-/-	-/15	-/-
D692/697 (OD1)	R638 (NH1)	-/-	-/-	-/-	-/-	-/13	-/-
D692/697 (OD1)	R638 (NH2)	-/-	-/-	-/-	-/-	-/37	-/-
D692/697 (OD2)	R638 (NE)	-/-	-/-	-/-	-/-	-/15	-/-
D692/697 (OD2)	R638 (NH1)	-/-	-/-	-/-	-/-	-/21	-/-
D692/697 (OD2)	R638 (NH2)	-/-	-/-	-/-	-/-	-/33	-/-
D692/697 (OD1)	K600 (NZ)	-/-	-/-	-/-	-/-	-/-	12/-
D692/697 (OD2)	K600 (NZ)	-/-	-/-	-/-	-/-	-/-	11/-
D692/697 (O)	K/M644 (NZ)	-/-	-/-	-/47	-/-	-/-	-/-
D692/697 (OD1)	K/M644 (NZ)	51/-	14/16	16/-	-/-	-/-	-/-
D692/697 (OD2)	K/M644 (NZ)	-/26	15/22	-/-	-/-	-/-	-/-
G693/698 (O)	N642 (N)	-/-	-/43	-/61	-/-	-/-	-/-
G693/698 (O)	N642 (ND2)	-/-	-/-	-/32	-/-	-/-	-/-
G693/698 (O)	K600 (NZ)	40/-	-/-	10/-	-/-	-/50	-/30
V695/700 (O)	K/M644 (N)	97/96	96/95	95/95	90/94	92/89	88/90

Parmi les résidus participant à la formation de liaison hydrogène, nous trouvons le résidu N642 dont la mutation N642H chez STAT5b a été répertoriée en clinique comme impliquée dans des formes agressives de leucémie à grands lymphocytes granuleux (*Large granular lymphocytes leukemia*, LGL) ou dans le développement de leucémies aigües à cellules T[365,366,400]. La mutation N642H se traduit *in vitro* par l'augmentation de l'activité de transcription et par l'augmentation de la phosphorylation de STAT5b[366,400]. Dans notre modèle théorique, N642 et V695/700 forment une liaison stable (*cf.* Figure 67). Néanmoins, cette liaison implique les atomes de la chaîne principale, des atomes conservés par la mutation. Ces données suggèrent donc un

effet de la chaîne latérale. Le changement de la chaîne latérale aurait un effet stabilisant sur la liaison entre monomères, potentiellement *via* la formation d'interactions avec le groupement phosphate du résidu pY699. En effet, le cycle imidazole est plus grand que la chaîne latérale de l'asparagine, et l'atome d'azote en position ε du résidu d'histidine pourrait se positionner à une distance permettant la formation d'interactions non covalentes. Cette hypothèse corrèle bien avec la présence variable (<5% du temps de simulation) de liaisons hydrogènes entre la chaine latérale de N642 et le groupement phosphate de Y699 observée dans les systèmes dSTAT5b^{HIE} de dSTAT5b^{HIP}. La modélisation et la simulation de ce mutant pourrait apporter des éléments de réponse pour déterminer les effets moléculaires de cette mutation.

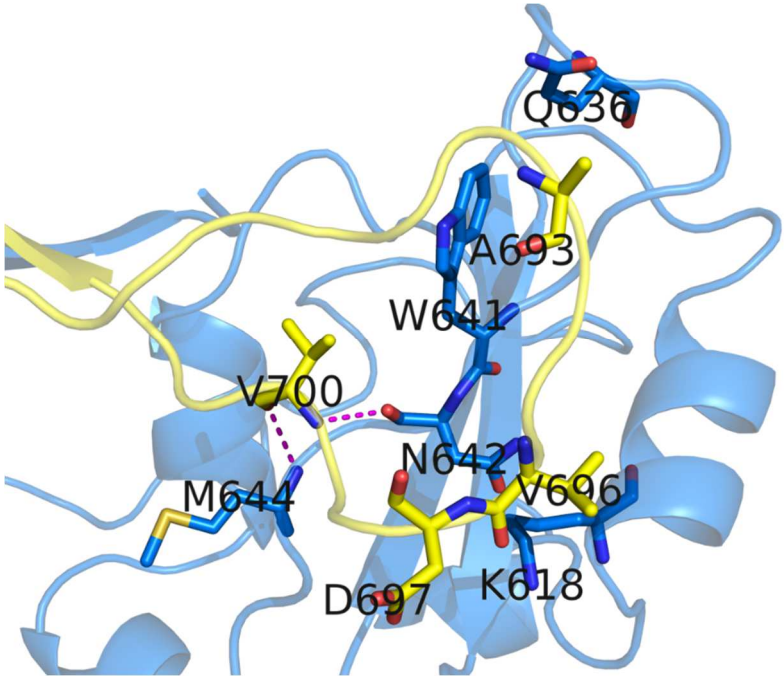


Figure 67 **: les interactions des dSTAT5.** Liaisons hydrogènes et résidus impliqués à proximité du résidu de phosphotyrosine (Y699). Les liaisons hydrogènes sont représentées en pointillés magenta. Un monomère est coloré en jaune, le second en bleu. Le système présenté est dSTAT5b^{HIP}.

Les autres mutants (T628S, T648S, R659C Y665F/H, I704L et Q706L) connus dans la littérature montrent des effets cliniques moins marqués que le mutant N642H. Le mutant T628S, dans le brin *B* du feuillet du domaine SH2, multiplie par 6 la transcription des gènes régulés par STAT5[405]. La mutation Y665F n'est ainsi pas associée avec des formes agressives de LGL[400], mais présente néanmoins une altération de la fonction de STAT5 caractérisée par une augmentation de l'activité transcriptionnelle et de la phosphorylation de STAT5b, à des niveaux toutefois inférieurs à ceux observés pour le mutant N642H. Le résidu Y665 se trouve sur la boucle reliant l'hélice αC à la queue (phospho-)tyrosyl, et n'est pas situé à l'interface STAT5 / STAT5. Le résidu T648 est situé au début de l'hélice αB, et n'est pas positionné à l'interface entre les monomères ou avec l'ADN. Le résidu R649 est néanmoins impliqué dans la liaison avec la chaîne principale de l'ADN (*cf.* Tableau 11). Le résidu R659 se trouve dans la boucle reliant les deux hélices *αB* et *αC*, et est impliqué dans une liaison hydrogène, mais entre les monomères A et B de dSTAT5a^{HID} (*cf.* Tableau 15), or le mutant R659C n'est observé que dans STAT5b. Les effets moléculaires sur

la stabilité des dimères de STAT5b des mutations de ces résidus restent cependant à caractériser. Enfin, les mutations des résidus I704 et Q706 sont à l'interface entre les deux queues phosphotyrosyl, et associée à une augmentation de la transcription[365,366,405].

F. Détection des ponts d'eau et cartes de densité du solvant dans dSTAT5

Les molécules d'eau situées aux interfaces des entités formant un complexe macromoléculaire peuvent constituer des éléments importants dans la stabilité des interactions et la flexibilité de ses deux molécules[551–553,617]. Afin de caractériser plus finement cette interface protéine – ADN, nous avons étudié les zones où des molécules d'eau sont fréquemment trouvées au cours de la simulation de dynamique moléculaire. Nous avons réalisé la superposition préalable de chaque conformation issue des dynamiques moléculaires en prenant comme référence les molécules d'ADN afin de ne pas prendre en compte les mouvements amples observés au niveau des domaines CCD.

L'analyse de toutes les simulations a montré des zones à l'interface protéine – ADN où les molécules d'eau ont une dynamique plus stable. Notamment, à l'intérieur des deux grands sillons de l'ADN dans lesquels STAT5 se fixe, on peut observer un espace volumineux occupé par des molécules d'eau de façon constante au cours des simulations de dynamique moléculaire (*cf.* Figure 68). Ces zones peuvent être connectées entre elles. D'autres zones moins grandes sont également observées en dehors du grand sillon, à l'interface entre la chaîne principale de l'ADN d'une part, et le feuillet *abe* et les hélices *α5* et *α8* d'autre part (*cf.* Figure 68).

Les zones mises en évidence et peuplées de molécules d'eau corrèlent bien avec la stabilité importante des régions protéiques adjacentes. Ces régions (feuillet *abe,* les hélices *α5* et *α8*) forment des contacts directs avec l'ADN, ce qui rigidifie cette structure et permet de présenter une interface stable. De nouveaux contacts peuvent ainsi se mettre en place entre ses domaines stables et l'ADN *via* les molécules d'eau, qui présentent habituellement une mobilité importante. Les interfaces protéines – ADN caractérisées par des mouvements plus importants (boucle entre les brins *e* et *f*) forment également des contacts avec l'ADN. Cependant, leur déplacement engendre une instabilité relative de l'interface avec l'ADN comparativement à l'interface du grand sillon. Ainsi, aucun volume peuplé d'eau n'est détecté dans ces régions.

La présence de ces regroupements de molécules d'eau suggère que des contacts protéine – ADN peuvent se faire *via* les molécules d'eau présentes à cette interface et jouent le rôle de pont d'eau. Afin de mieux caractériser ces contacts, nous avons développé un script d'analyse des liaisons hydrogènes entre les molécules impliquées dans ces interactions à l'interface des différentes entités moléculaires, à savoir les deux monomères de STAT5, le double brin d'ADN et les molécules d'eau. Dans le cas présent, nous avons étudié d'une part les liaisons hydrogènes entre STAT5 et le solvant, puis entre l'ADN et l'eau. Nous avons ensuite croisé ces résultats afin de détecter les ponts d'eau. Ces structures consistent en une liaison transmise par une molécule d'eau entre deux molécules protéiques. Soit les deux molécules A et B (A, B =

protéine ou ADN) et la molécule d'eau C, si A et B forment une liaison hydrogène avec C au même moment, un pont d'eau est établi entre A et B.

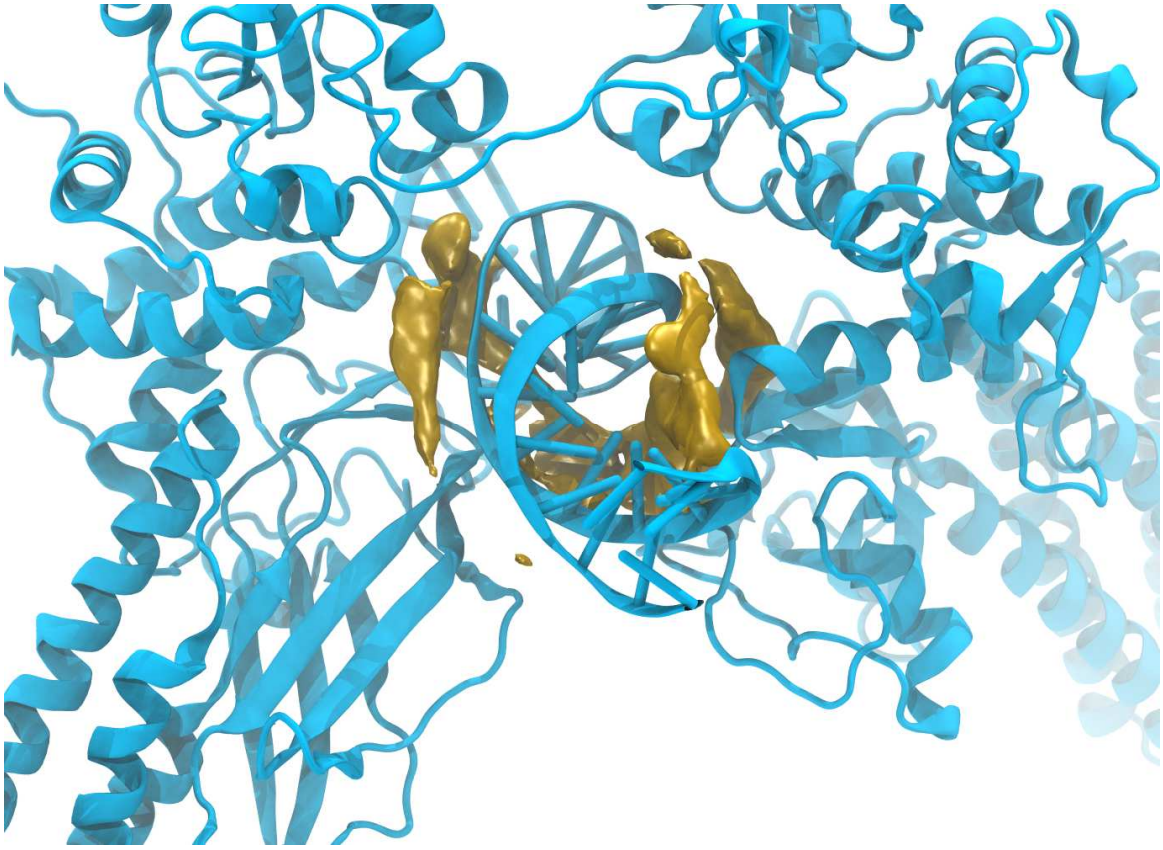


Figure 68 : **Zones protéine - ADN stabilisant les molécules d'eau pour dSTAT5b^{HID}.** Les volumes jaunes indiquent les zones occupées par l'eau de manière constante. La conformation affichée est la conformation moyenne.

1. *Description de la structure du script de détection des ponts d'eau*

Le script de détection des ponts d'eau se base sur des fichiers de sortie optionnels de la fonction *gromacs g_hbond*. Ces fichiers sont d'une part le '*fichier index*' contenant les numéros des atomes formant les liaisons hydrogènes (produit avec la commande *hbn*), et d'autre part le '*fichier matrice*' correspondant qui indique les conformations dans lesquelles chaque liaison hydrogène est retrouvée. En fonction des groupes moléculaire entre lesquelles les liaisons hydrogènes sont calculées et du nombre de conformations analysées, le second fichier peut être très volumineux (de l'ordre de plusieurs gigaoctets, Go). L'analyse des ponts d'eau se place typiquement dans cette catégorie de fichier, les molécules d'eau étant nombreuses et possédant la capacité de former des liaisons hydrogènes facilement. La molécule d'eau est en effet constituée d'un atome d'oxygène qui peut être accepteur de liaisons hydrogènes mais également donneur de liaisons par ses deux atomes d'hydrogène. Au cours d'une simulation de dynamique moléculaire, plusieurs millions de liaisons impliquant des molécules d'eau peuvent ainsi se former.

Le script détaillé ci-dessous prend en entrée deux paires de fichiers accompagnées d'un fichier PDB standard. Chaque paire de fichiers est composée d'un '*fichier index*' et du '*fichier matrice*' correspondant, permettant ainsi de décrire l'ensemble des liaisons hydrogènes observées au cours de la trajectoire entre deux groupes de molécules. L'un des groupes est constitué des molécules d'eau de la boîte de simulation, l'autre regroupe des résidus protéiques ou nucléotidiques. Les ponts d'eau vont être donc détectés entre les deux groupes constitués de résidus de nucléotides ou de protéine. Les sorties générées par le script pour chaque paire de fichier récapitulent les résidus et atomes donneurs et accepteurs de chaque liaison hydrogène, ainsi que l'occurrence, exprimée en pourcentage du temps de simulation, de survenue de la liaison. Pour les ponts d'eau, deux fichiers sont générés : un '*fichier matrice*' qui reprend le même format que les fichiers matrice générés par *gromacs*, et un '*fichier texte*' récapitulant les résidus qui forment chaque pont d'eau, et son occurrence exprimée en pourcentage du temps de simulation.

Le choix de cette structure de fichier a été inspiré par le script plot_hbmap.pl créé et mis à disposition par J. Lemkul à l'adresse suivante : http://www.bevanlab.biochem.vt.edu/Pages/Personal/justin/scripts.html . Ce script a été écrit afin de calculer la fraction de temps pendant laquelle chaque liaison hydrogène existe. Nous avons réutilisé et adapté ce script dans un premier temps pour calculer les liaisons hydrogènes de chaque paire de fichiers, puis développé un algorithme permettant de confronter ces deux résultats l'un à l'autre pour calculer les ponts d'eau (*cf.* Figure 69 et Annexe B.).

La première fonction (hbmapPreProc) permet de prétraiter les fichiers d'entrée et d'initialiser les structures de données qui seront remplies à l'étape suivante. La fonction est appelée deux fois successivement pour traiter les deux paires de fichier '*index*' et '*matrice*' selon le protocole suivant:

Les variables destinées à être remplies par le nom des résidus et des atomes donneurs et accepteurs des liaisons hydrogènes pour chaque paire de fichier sont initialisées. La fonction *hbmapCoord* est parallélisée par l'emploi du module *threads*, qui permet au processus parent (le script) de créer plusieurs processus enfants qui vont traiter simultanément plusieurs tâches. Les variables initialisées sont partagées pour que chaque processus enfant ait accès à ces données. Chaque processus enfant va traiter une ligne du fichier *pdb* d'entrée et remplit les variables partagées.

Une fois remplies, ces variables partagées et les fichiers '*matrice*' correspondants sont parcourus pour imprimer les fichiers récapitulant les liaisons hydrogènes entre chaque molécule et l'eau. Cette étape est réalisée par la fonction *hbmapOutput* (lignes 467 – 495, *cf.* Annexe B.).

Le script *plot_hbmap.pl* de J. Lemkul s'arrête à cette étape. Nous l'avons déjà adapté pour qu'il traite les deux paires de fichier, mais également pour que l'étape la plus couteuse en temps de calcul soit parallélisable sur plusieurs processeurs (lignes 178 – 199, *cf.* Annexe B.). La suite du script a été développée au cours de la thèse.

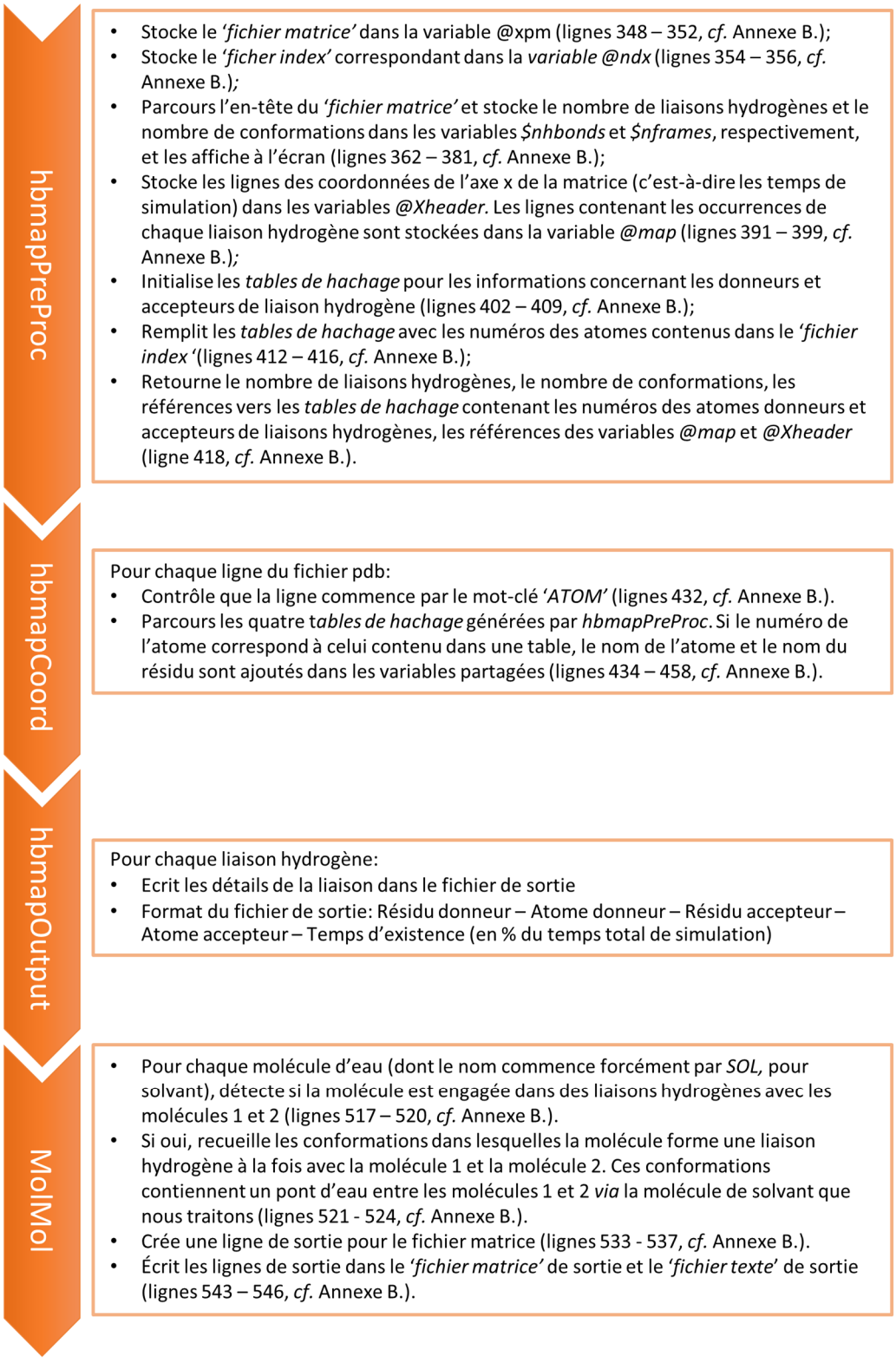


Figure 69 : **Flux des fontions successives appelées pour détecter les ponts d'eau.** Pour chaque fontion, une courte description des opérations réalisées est affichée dans l'encadré correspondant.

L'étape suivante écrit l'en-tête du '*fichier matrice*' de sortie, et ouvre le '*fichier texte*' de sortie. La fonction *MolMol*, qui détecte les ponts d'eau, est appelée quatre fois successivement afin de traiter les quatre types de ponts d'eau possible (lignes 243 – 330, *cf.* Annexe B.) : (1) les

molécules d'eau sont les donneurs de liaisons hydrogènes, (2) les molécules d'eau sont donneurs de liaison hydrogène vers la molécule 1 et accepteurs de liaisons hydrogènes de la molécule 2, (3) les molécules d'eau sont donneurs de liaison hydrogène vers la molécule 2 et accepteurs de liaisons hydrogènes de la molécule 1, (4) les molécules d'eau sont accepteurs de liaisons hydrogènes.

Finalement, les fichiers de sortie contiennent les informations voulues et peuvent ensuite être analysés en fonction des besoins. Plusieurs points importants sont à noter néanmoins.

Tout d'abord, les temps de calcul nécessaires sont dépendants du nombre de liaisons hydrogènes entre chaque molécule et l'eau. Ainsi, et afin de réduire le temps de calcul, nous avons limité l'analyse des liaisons hydrogènes initiales aux domaines DBD pour la protéine. Malgré la parallélisation de calcul des étapes les plus longues du script, le facteur limitant reste le volume de mémoire nécessaire. En effet, les données sont toutes stockées dans la mémoire, or les fichiers d'entrée sont constitués de plusieurs fichiers pouvant atteindre plusieurs giga-octets (Go). La consommation de mémoire vive est donc conséquente. De plus, chaque processus '*enfant*' de la fonction *MolMol* duplique les variables afin de pouvoir les traiter, ce qui duplique également l'espace de mémoire vive nécessaire au lancement du script. Plusieurs dizaines de Go sont donc requis pour lancer ce script, ce qui n'est pas compatible avec une utilisation en routine, sur un ordinateur de bureau. Les prochaines étapes de développement vont s'attacher à diminuer ces exigences. Pour cela, plusieurs pistes sont à l'étude. D'une part, l'optimisation de ce script et l'adoption de structures de données plus adaptées sont considérées. L'emploi de modules permettant de travailler sur les données sans les placer dans la mémoire vive est également une piste intéressante. Le partage des données entre les processus '*enfants*' devrait également limiter les besoins en mémoire vive. Enfin, le portage vers un langage plus adapté tel que C/C++ pourrait constituer une alternative intéressante car le contrôle de la mémoire est plus aisé.

2. Ponts d'eau à l'interface protéine - ADN

L'emploi de ce script nous a permis de détecter de nombreux ponts d'eau chez dSTAT5a^{HID}. En raison de la longueur des calculs, les ponts d'eau dans les autres systèmes n'ont pas été caractérisés.

La grande majorité (98%) des ponts d'eau détectés ne sont présents qu'au cours d'un petit nombre de conformations. Les occurrences des ponts d'eau sont en conséquence très faibles (< 5% du temps de simulation, soit 1,5 ns. Cependant, les ponts d'eau sont constitués d'un *triplet* résidu A – molécule d'eau C– nucléotide B spécifique, donc plusieurs ponts d'eau entre le résidu A et le nucléotide B peuvent exister successivement. Le développement d'une

fonction additionnelle est envisagé pour détecter de telles structures et permettre une caractérisation plus fine des ponts d'eau.

Les ponts d'eau aqueux sont observés indifféremment entre la chaîne principale ou les chaînes latérales des résidus, et les bases azotées ou la partie désoxyribose des nucléotides. Par exemple, un pont d'eau est observé entre la fonction carboxylique de la chaîne latérale du résidu Q378et la base azotée de l'adénine en position +6 du double brin d'ADN dans plus de 3,7% des conformations. La multiplication de ce type d'interactions au cours de la dynamique pourrait ainsi constituer un vecteur important de l'affinité et de la spécificité de reconnaissance entre STAT5 et l'ADN. Ainsi, plus de 16 800 ponts d'eau individuels sont détectés au cours de 30 ns de simulations de dynamique moléculaire, alors que seulement 306 interactions directes entre STAT5 et l'ADN sont observées. Cependant, les liaisons formées par les ponts d'eau ont un caractère beaucoup plus instable que les liaisons hydrogènes directes, ce qui rend la comparaison entre ces deux valeurs incertaine. L'analyse détaillée de ces interactions permettra d'affiner la connaissance de l'interface STAT5 / ADN, et de déterminer le rôle joué par les molécules d'eau, notamment en comparant ce type d'interaction aux liaisons hydrogènes directes.

Conclusion et perspectives

STAT5 (*Signal Transducer and Activator of Transcription 5*) est une protéine clé dans la transmission de nombreux signaux intracellulaires, régulant ainsi de nombreuses fonctions physiologiques, en particulier au cours de l'hématopoïèse. L'activation de STAT5 nécessite sa phosphorylation sur résidu tyrosine par des protéines partenaires (dont les kinases Janus). STAT5 se dimérise alors et peut alors passer dans le noyau cellulaire où elle favorise la transcription de gènes clés après s'être fixé sur des séquences cibles. STAT5 régule ainsi l'expression de protéines impliquées dans la progression du cycle cellulaire, la prolifération et la survie cellulaire. La dérégulation des activateurs de STAT5 vers des formes constitutivement actives entraîne l'augmentation de la phosphorylation de STAT5 donc de son activité, et la transmission d'un signal oncogénique. STAT5 est ainsi impliquée dans la physiopathologie de multiples cancers, dont la leucémie myéloïde chronique et les mastocytoses. D'autre part, des études récentes ont fait part de mutations activatrices acquises de l'isoforme STAT5b impliquées dans la physiopathologie de néoplasies myéloprolifératives. De nombreuses études ont par ailleurs montré que STAT5 est nécessaire au développement des tumeurs et à la survie des cellules cancéreuses, notamment dans le cas de tumeurs résistantes aux traitements actuels. Toutes ces données suggèrent donc que STAT5 est une cible pertinente pour le développement de thérapies innovantes dans le domaine de l'onco-hématologie.

La caractérisation de structures (par cristallographie des rayons X ou RMN) permet l'élaboration de modèles structuraux à l'échelle atomique, étape nécessaire à l'étude des phénomènes dynamiques au sein des complexes macromoléculaires. Les méthodes bio-informatiques associées aux données de la biologie structurale permettent ainsi l'exploration de la relation séquence – structure – dynamique – fonction des protéines *via* des approches de biologie computationnelle innovantes. Les simulations de dynamique moléculaire (DM) explorent l'espace conformationnel des macromolécules biologiques à l'échelle atomique, ce qui permet de compléter les observations expérimentales. L'analyse des trajectoires de DM des monomères des deux isoformes de STAT5 a mis en évidence les propriétés dynamiques de la partie distale du CCD, ainsi que la présence de variations structurales liées aux différences de la séquence primaire de STAT5a et STAT5b. L'impact de la phosphorylation sur la structure et la dynamique de STAT5 a également été caractérisé dans nos travaux. Le couplage dynamique longue-distance décrit par deux méthodes analytiques indépendantes a été étudié *via* l'analyse du réseau de communication intra-protéine, en utilisant la méthode MONETA, développée au sein de notre équipe pour analyser les effets allostériques. La présence des mouvements amples de STAT5 empêchant l'application du MONETA a motivé le développement d'une méthode originale pour décomposer la dynamique de STAT5 à l'échelle global et local, et détecter des caractéristiques de la dynamique moléculaire locale. La collaboration avec Pr. A. Trouvé (CMLA, ENS Cachan) a permis l'implémentation de l'algorithme de « décomposition des traits

principaux, PFD » dans MONETA. Par cette nouvelle approche appliquée au traitement de la dynamique des monomères de STAT5, nous avons pu montrer une grande similarité des mouvements locaux mais aussi des particularités dynamiques liées à la séquence et/ou à la présence du groupement phosphate. L'analyse comparative des chemins de communication à travers STAT5 phosphorylé ou non-phosphorylé a révélé l'impact longue distance de la phosphorylation du résidu tyrosine qui perturbe le réseau des chemins de communication. Par ailleurs, la recherche de poches à la surface des protéines STAT5 nous a permis de localiser et caractériser une poche proche du site de fixation de la phosphotyrosine, ainsi qu'une seconde poche adjacente. La description des chemins de communication à travers STAT5 a mis en évidence que ces deux poches sont localisées à proximité des voies de communications des STAT5 analysés. Les divergences de la communication intramoléculaire entre les deux isoformes de STAT5 et du comportement dynamique des poches permettent de considérer ces poches comme des sites allostériques de fixation de molécules capables de moduler les fonctions de STAT5. . Ces poches pourront être exploitées pour développer des modulateurs de l'activité de STAT5 et offrent de nouvelles pistes pour le développement de modulateurs spécifiques. D'autre part, les trajectoires de MD fournissent des données cruciales – structurales dans le cadre d'un projet de criblage virtuel d'une chimiothèque, et dynamiques dans l'optique de compréhension de la régulation allostérique de cette famille de protéines.

Les simulations de dynamique moléculaire des dimères de STAT5 liés à un fragment d'ADN ont révélé les mouvements amples des domaines CCD de chaque monomère, similaires à ceux d'une paire de ciseaux. Un mouvement similaire a été publié pour STAT3, et suggère ainsi que les dimères de la famille des STATs partagent ce type mouvement. L'analyse des interactions à l'interface entre les monomères STAT5 a mis en évidence la similarité du site de liaison de la phosphotyrosine dans les deux isoformes STAT5a et STAT5b, et confirmé le rôle clé joué par les résidus conservés dans toutes les protéines STATs. Les données issues des trajectoires de DM nous ont également permis d'étudier le rôle des résidus mutés dans plusieurs cohortes de patients atteints de leucémie. Le résidu N642, impliqué dans des formes agressives de leucémies à grands lymphocytes granuleux, est localisé en périphérie du site de fixation de la phosphotyrosine et sa mutation pourrait permettre la formation de nouvelles interactions entre les monomères, stabilisant le dimère lié à l'ADN. Cette hypothèse semble en accord avec les données expérimentales obtenues *in vitro* montrant une augmentation de l'activité de transcription du mutant N642H de STAT5. Les mutations des résidus I704 et Q706 est également retrouvée chez des patients atteints de leucémies, et sont associées à une augmentation modérée de la transcription. Ces résidus sont situés à l'interface entre les deux extrémités C-terminale du *Core Fragment*, leurs mutations pourraient donc créer de nouvelles interactions et stabiliser le complexe STAT5 - ADN. Les autres résidus mutés sont à la surface de STAT5, en dehors des régions impliquées dans la formation du dimère. Ces mutants impliquent donc d'autres effets, comme par exemple la liaison à des protéines partenaires menant à la stabilisation des complexes de transcription. La construction d'un complexe composé de STAT5 et d'une protéine activatrice permettrait l'étude approfondie de cette hypothèse. L'étude de l'interface protéine – ADN au sein des dimères de STAT5 nous a permis de montrer que la

protonation du résidu d'histidine 471 joue un rôle dans la reconnaissance de la molécule d'ADN par STAT5 *via* la formation de liaisons hydrogènes avec les bases azotées. Les dimères présentant la double protonation du résidu H471 sont ainsi corrélés à la formation de liaisons hydrogène spécifiques. L'analyse de l'énergie d'interaction STAT5 – ADN en fonction de l'état de protonation de H471 complèterait la description de cette interface. Enfin, nous avons montré que les molécules d'eau forment des ponts d'eau reliant STAT5 à l'ADN, apportant une interface supplémentaire. L'exploration détaillée du rôle des molécules d'eau dans la reconnaissance à l'ADN par STAT5 et l'affinité de cette liaison permettraient d'expliquer les variations de leur affinité en fonction de la séquence d'ADN.

L'ensemble de nos résultats constituent la première caractérisation de la structure et de la dynamique des différentes formes de STAT5 à l'échelle atomique. Nous avons pu établir des liens entre les changements dynamiques observés par DM et les variations de la séquence primaire des isoformes de STAT5 monomériques, et montré des différences du réseau allostérique à proximité de sites de liaison potentiels de modulateurs. Ces données pourraient donc ouvrir la voie et offrir une stratégie de développement de modulateurs innovants de l'activité de STAT5. Enfin, l'étude des formes dimériques liées à l'ADN est en accord avec les données expérimentales, et apporte des éléments nouveaux d'une part dans la description des mécanismes de la reconnaissance STAT5 – ADN et STAT5 – STAT5, et d'autre part dans l'étude des mutants de STAT5b observées en clinique. La poursuite de ces travaux pourrait aboutir à la description des phénomènes moléculaires expliquant les différences d'activité de STAT5 et de ses mutants.

Références bibliographiques

1. Yan, R., Qureshi, S., Zhong, Z., Wen, Z. & Darnell, J. E. The genomic structure of the STAT genes: multiple exons in coincident sites in Stat1 and Stat2. *Nucleic Acids Res.* **23,** 459–463 (1995).
2. Yamamoto, K. *et al.* Stat4, a novel gamma interferon activation site-binding protein expressed in early myeloid differentiation. *Mol. Cell. Biol.* **14,** 4342–4349 (1994).
3. Thierfelder, W. E. *et al.* Requirement for Stat4 in interleukin-12-mediated responses of natural killer and T cells. *Nature* **382,** 171–174 (1996).
4. Hou, J. *et al.* An interleukin-4-induced transcription factor: IL-4 Stat. *Science* **265,** 1701–1706 (1994).
5. Leek, J. P., Hamlin, P. J., Bell, S. M. & Lench, N. J. Assignment of the STAT6 gene (STAT6) to human chromosome band 12q13 by in situ hybridization. *Cytogenet. Cell Genet.* **79,** 208–209 (1997).
6. Akira, S. *et al.* Molecular cloning of APRF, a novel IFN-stimulated gene factor 3 p91-related transcription factor involved in the gp130-mediated signaling pathway. *Cell* **77,** 63–71 (1994).
7. Hou, J., Schindler, U., Henzel, W. J., Wong, S. C. & McKnight, S. L. Identification and purification of human Stat proteins activated in response to interleukin-2. *Immunity* **2,** 321–329 (1995).
8. Lin, J. X., Mietz, J., Modi, W. S., John, S. & Leonard, W. J. Cloning of human Stat5B. Reconstitution of interleukin-2-induced Stat5A and Stat5B DNA binding activity in COS-7 cells. *J. Biol. Chem.* **271,** 10738–10744 (1996).
9. Copeland, N. G. *et al.* Distribution of the mammalian Stat gene family in mouse chromosomes. *Genomics* **29,** 225–228 (1995).
10. Wang, D. A small amphipathic alpha -helical region is required for transcriptional activities and proteasome-dependent turnover of the tyrosine-phosphorylated Stat5. *EMBO J.* **19,** 392–399 (2000).
11. Darnell, J. E., Kerr, I. M. & Stark, G. R. Jak-STAT pathways and transcriptional activation in response to IFNs and other extracellular signaling proteins. *Science* **264,** 1415–1421 (1994).
12. Bibi, S. *et al.* Co-operating STAT5 and AKT signaling pathways in chronic myeloid leukemia and mastocytosis: possible new targets of therapy. *Haematologica* **99,** 417–429 (2014).
13. Schindler, C., Shuai, K., Prezioso, V. R. & Darnell, J. E. Interferon-dependent tyrosine phosphorylation of a latent cytoplasmic transcription factor. *Science* **257,** 809–813 (1992).
14. Shuai, K., Schindler, C., Prezioso, V. R. & Darnell, J. E. Activation of transcription by IFN-gamma: tyrosine phosphorylation of a 91-kD DNA binding protein. *Science* **258,** 1808–1812 (1992).
15. Shuai, K., Stark, G. R., Kerr, I. M. & Darnell, J. E. A single phosphotyrosine residue of Stat91 required for gene activation by interferon-gamma. *Science* **261,** 1744–1746 (1993).
16. Velazquez, L., Fellous, M., Stark, G. R. & Pellegrini, S. A protein tyrosine kinase in the interferon αβ signaling pathway. *Cell* **70,** 313–322 (1992).
17. Müller, M. *et al.* The protein tyrosine kinase JAK1 complements defects in interferon-alpha/beta and -gamma signal transduction. *Nature* **366,** 129–135 (1993).
18. Garcia, R. *et al.* Constitutive activation of Stat3 by the Src and JAK tyrosine kinases participates in growth regulation of human breast carcinoma cells. *Oncogene* **20,** 2499–2513 (2001).
19. Vignais, M. L., Sadowski, H. B., Watling, D., Rogers, N. C. & Gilman, M. Platelet-derived growth factor induces phosphorylation of multiple JAK family kinases and STAT proteins. *Mol. Cell. Biol.* **16,** 1759–1769 (1996).

20. Legeai-Mallet, L., Benoist-Lasselin, C., Delezoide, A.-L., Munnich, A. & Bonaventure, J. Fibroblast Growth Factor Receptor 3 Mutations Promote Apoptosis but Do Not Alter Chondrocyte Proliferation in Thanatophoric Dysplasia. *J. Biol. Chem.* **273,** 13007–13014 (1998).
21. Ferrand, A. *et al.* A novel mechanism for JAK2 activation by a G protein-coupled receptor, the CCK2R: implication of this signaling pathway in pancreatic tumor models. *J. Biol. Chem.* **280,** 10710–10715 (2005).
22. Wong, M. & Fish, E. N. RANTES and MIP-1alpha activate stats in T cells. *J. Biol. Chem.* **273,** 309–314 (1998).
23. Danial, N. N. *et al.* Direct interaction of Jak1 and v-Abl is required for v-Abl-induced activation of STATs and proliferation. *Mol. Cell. Biol.* **18,** 6795–6804 (1998).
24. Gao, X., Wang, H., Yang, J. J., Liu, X. & Liu, Z.-R. Pyruvate kinase M2 regulates gene transcription by acting as a protein kinase. *Mol. Cell* **45,** 598–609 (2012).
25. Hanson, E. M., Dickensheets, H., Qu, C.-K., Donnelly, R. P. & Keegan, A. D. Regulation of the dephosphorylation of Stat6. Participation of Tyr-713 in the interleukin-4 receptor alpha, the tyrosine phosphatase SHP-1, and the proteasome. *J. Biol. Chem.* **278,** 3903–3911 (2003).
26. Johnson, D. J. *et al.* Shp1 regulates T cell homeostasis by limiting IL-4 signals. *J. Exp. Med.* **210,** 1419–1431 (2013).
27. Lu, X. *et al.* T-Cell Protein Tyrosine Phosphatase, Distinctively Expressed in Activated-B-Cell-Like Diffuse Large B-Cell Lymphomas, Is the Nuclear Phosphatase of STAT6. *Mol. Cell. Biol.* **27,** 2166–2179 (2007).
28. Lu, X. *et al.* PTP1B is a negative regulator of interleukin 4-induced STAT6 signaling. *Blood* **112,** 4098–4108 (2008).
29. Nakahira, M., Tanaka, T., Robson, B. E., Mizgerd, J. P. & Grusby, M. J. Regulation of Signal Transducer and Activator of Transcription Signaling by the Tyrosine Phosphatase PTP-BL. *Immunity* **26,** 163–176 (2007).
30. Böhmer, F.-D. & Friedrich, K. Protein tyrosine phosphatases as wardens of STAT signaling. *JAK-STAT* **3,** e28087 (2014).
31. Ketteler, R. *et al.* The cytokine-inducible Scr homology domain-containing protein negatively regulates signaling by promoting apoptosis in erythroid progenitor cells. *J. Biol. Chem.* **278,** 2654–2660 (2003).
32. Narazaki, M. *et al.* Three distinct domains of SSI-1/SOCS-1/JAB protein are required for its suppression of interleukin 6 signaling. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 13130–13134 (1998).
33. Favre, H., Benhamou, A., Finidori, J., Kelly, P. A. & Edery, M. Dual effects of suppressor of cytokine signaling (SOCS-2) on growth hormone signal transduction. *FEBS Lett.* **453,** 63–66 (1999).
34. Greenhalgh, C. J. *et al.* Biological Evidence That SOCS-2 Can Act Either as an Enhancer or Suppressor of Growth Hormone Signaling. *J. Biol. Chem.* **277,** 40181–40184 (2002).
35. Cacalano, N. A., Sanden, D. & Johnston, J. A. Tyrosine-phosphorylated SOCS-3 inhibits STAT activation but binds to p120 RasGAP and activates Ras. *Nat. Cell Biol.* **3,** 460–465 (2001).
36. Seki, Y. -i. *et al.* Expression of the suppressor of cytokine signaling-5 (SOCS5) negatively regulates IL-4-dependent STAT6 activation and Th2 differentiation. *Proc. Natl. Acad. Sci.* **99,** 13003–13008 (2002).
37. Lim, C. P. & Cao, X. Structure, function, and regulation of STAT proteins. *Mol.Biosyst.* **2,** 536–550 (2006).
38. Liu, B. *et al.* Inhibition of Stat1-mediated gene activation by PIAS1. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 10626–10631 (1998).
39. Kotaja, N., Karvonen, U., Jänne, O. A. & Palvimo, J. J. PIAS proteins modulate transcription factors by functioning as SUMO-1 ligases. *Mol. Cell. Biol.* **22,** 5222–5234 (2002).

40. Wormald, S. & Hilton, D. J. Inhibitors of cytokine signal transduction. *J. Biol. Chem.* **279,** 821–824 (2004).
41. Arora, T. *et al.* PIASx is a transcriptional co-repressor of signal transducer and activator of transcription 4. *J. Biol. Chem.* **278,** 21327–21330 (2003).
42. Sachdev, S. *et al.* PIASy, a nuclear matrix-associated SUMO E3 ligase, represses LEF1 activity by sequestration into nuclear bodies. *Genes Dev.* **15,** 3088–3103 (2001).
43. Hoey, T. *et al.* Distinct requirements for the naturally occurring splice forms Stat4alpha and Stat4beta in IL-12 responses. *EMBO J.* **22,** 4237–4248 (2003).
44. Schaefer, T. S., Sanders, L. K. & Nathans, D. Cooperative transcriptional activity of Jun and Stat3 beta, a short form of Stat3. *Proc. Natl. Acad. Sci. U. S. A.* **92,** 9097–9101 (1995).
45. Schindler, C., Fu, X. Y., Improta, T., Aebersold, R. & Darnell, J. E. Proteins of transcription factor ISGF-3: one gene encodes the 91-and 84-kDa ISGF-3 proteins that are activated by interferon alpha. *Proc. Natl. Acad. Sci.* **89,** 7836–7839 (1992).
46. Wang, D., Stravopodis, D., Teglund, S., Kitazawa, J. & Ihle, J. N. Naturally occurring dominant negative variants of Stat5. *Mol. Cell. Biol.* **16,** 6141–6148 (1996).
47. Patel, B. K., Pierce, J. H. & LaRochelle, W. J. Regulation of interleukin 4-mediated signaling by naturally occurring dominant negative and attenuated forms of human Stat6. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 172–177 (1998).
48. Caldenhoven, E. *et al.* STAT3beta, a splice variant of transcription factor STAT3, is a dominant negative regulator of transcription. *J. Biol. Chem.* **271,** 13221–13227 (1996).
49. Maritano, D. *et al.* The STAT3 isoforms alpha and beta have unique and specific functions. *Nat. Immunol.* **5,** 401–409 (2004).
50. Müller, M. *et al.* Complementation of a mutant cell line: central role of the 91 kDa polypeptide of ISGF3 in the interferon-alpha and -gamma signal transduction pathways. *EMBO J.* **12,** 4221–4228 (1993).
51. Azam, M. *et al.* Interleukin-3 signals through multiple isoforms of Stat5. *EMBO J.* **14,** 1402–1411 (1995).
52. Azam, M., Lee, C., Strehlow, I. & Schindler, C. Functionally distinct isoforms of STAT5 are generated by protein processing. *Immunity* **6,** 691–701 (1997).
53. Chakraborty, A. & Tweardy, D. J. Granulocyte colony-stimulating factor activates a 72-kDa isoform of STAT3 in human neutrophils. *J. Leukoc. Biol.* **64,** 675–680 (1998).
54. Darnowski, J. W. *et al.* Stat3 cleavage by caspases: impact on full-length Stat3 expression, fragment formation, and transcriptional activity. *J. Biol. Chem.* **281,** 17707–17717 (2006).
55. Hevehan, D. L., Miller, W. M. & Papoutsakis, E. T. Differential expression and phosphorylation of distinct STAT3 proteins during granulocytic differentiation. *Blood* **99,** 1627–1637 (2002).
56. Lee, C. *et al.* Characterization of the Stat5 protease. *J. Biol. Chem.* **274,** 26767–26775 (1999).
57. Meyer, J., Jucker, M., Ostertag, W. & Stocking, C. Carboxyl-truncated STAT5beta is generated by a nucleus-associated serine protease in early hematopoietic progenitors. *Blood* **91,** 1901–1908 (1998).
58. Mitchell, T. J., Whittaker, S. J. & John, S. Dysregulated expression of COOH-terminally truncated Stat5 and loss of IL2-inducible Stat5-dependent gene expression in Sezary Syndrome. *Cancer Res.* **63,** 9048–9054 (2003).
59. Sherman, M. A., Secor, V. H. & Brown, M. A. IL-4 preferentially activates a novel STAT6 isoform in mast cells. *J. Immunol. Baltim. Md 1950* **162,** 2703–2708 (1999).
60. Suzuki, K. *et al.* Stat6-protease but not Stat5-protease is inhibited by an elastase inhibitor ONO-5046. *Biochem. Biophys. Res. Commun.* **309,** 768–773 (2003).
61. Xia, Z. *et al.* A novel serine-dependent proteolytic activity is responsible for truncated signal transducer and activator of transcription proteins in acute myeloid leukemia blasts. *Cancer Res.* **61,** 1747–1753 (2001).

62. tenOever, B. R. *et al.* Multiple Functions of the IKK-Related Kinase IKK in Interferon-Mediated Antiviral Immunity. *Science* **315,** 1274–1278 (2007).
63. Beuvink, I. *et al.* Stat5a serine phosphorylation. Serine 779 is constitutively phosphorylated in the mammary gland, and serine 725 phosphorylation influences prolactin-stimulated in vitro DNA binding activity. *J. Biol. Chem.* **275,** 10247–10255 (2000).
64. Clark, D. E. *et al.* ERBB4/HER4 Potentiates STAT5A Transcriptional Activity by Regulating Novel STAT5A Serine Phosphorylation Events. *J. Biol. Chem.* **280,** 24175–24180 (2005).
65. Steen, H. C. *et al.* Identification of STAT2 Serine 287 as a Novel Regulatory Phosphorylation Site in Type I Interferon-induced Cellular Responses. *J. Biol. Chem.* **288,** 747–758 (2013).
66. Yuan, Z. -l. Stat3 Dimerization Regulated by Reversible Acetylation of a Single Lysine Residue. *Science* **307,** 269–273 (2005).
67. Kramer, O. H. *et al.* A phosphorylation-acetylation switch regulates STAT1 signaling. *Genes Dev.* **23,** 223–235 (2009).
68. Wieczorek, M., Ginter, T., Brand, P., Heinzel, T. & Krämer, O. H. Acetylation modulates the STAT signaling code. *Cytokine Growth Factor Rev.* **23,** 293–305 (2012).
69. Mowen, K. A. *et al.* Arginine methylation of STAT1 modulates IFNalpha/beta-induced transcription. *Cell* **104,** 731–741 (2001).
70. Chen, W., Daines, M. O. & Hershey, G. K. K. Methylation of STAT6 Modulates STAT6 Phosphorylation, Nuclear Translocation, and DNA-Binding Activity. *J. Immunol.* **172,** 6744–6750 (2004).
71. Duong, F. H. T., Filipowicz, M., Tripodi, M., La Monica, N. & Heim, M. H. Hepatitis C virus inhibits interferon signaling through up-regulation of protein phosphatase 2A. *Gastroenterology* **126,** 263–277 (2004).
72. Rho, J., Choi, S., Seong, Y. R., Choi, J. & Im, D.-S. The Arginine-1493 Residue in QRRGRTGR1493G Motif IV of the Hepatitis C Virus NS3 Helicase Domain Is Essential for NS3 Protein Methylation by the Protein Arginine Methyltransferase 1. *J. Virol.* **75,** 8031–8044 (2001).
73. Komyod, W., Bauer, U.-M., Heinrich, P. C., Haan, S. & Behrmann, I. Are STATS Arginine-methylated? *J. Biol. Chem.* **280,** 21700–21705 (2005).
74. Meissner, T., Krause, E., Lödige, I. & Vinkemeier, U. Arginine Methylation of STAT1. *Cell* **119,** 587–589 (2004).
75. Mowen, K. & David, M. Response to Matters Arising. *Cell* **119,** 589–590 (2004).
76. Iwasaki, H. *et al.* Disruption of Protein Arginine N-Methyltransferase 2 Regulates Leptin Signaling and Produces Leanness In Vivo Through Loss of STAT3 Methylation. *Circ. Res.* **107,** 992–1001 (2010).
77. Yang, J. *et al.* Reversible methylation of promoter-bound STAT3 by histone-modifying enzymes. *Proc. Natl. Acad. Sci.* **107,** 21499–21504 (2010).
78. Kim, E. *et al.* Phosphorylation of EZH2 Activates STAT3 Signaling via STAT3 Methylation and Promotes Tumorigenicity of Glioblastoma Stem-like Cells. *Cancer Cell* **23,** 839–852 (2013).
79. Kim, T. K. & Maniatis, T. Regulation of Interferon-gamma -Activated STAT1 by the Ubiquitin-Proteasome Pathway. *Science* **273,** 1717–1719 (1996).
80. Elliott, J. *et al.* Respiratory Syncytial Virus NS1 Protein Degrades STAT2 by Using the Elongin-Cullin E3 Ligase. *J. Virol.* **81,** 3428–3436 (2007).
81. Wei, J. *et al.* The Ubiquitin Ligase TRAF6 Negatively Regulates the JAK-STAT Signaling Pathway by Binding to STAT3 and Mediating Its Ubiquitination. *PLoS ONE* **7,** e49567 (2012).
82. Rogers, R. S., Horvath, C. M. & Matunis, M. J. SUMO Modification of STAT1 and Its Role in PIAS-mediated Inhibition of Gene Activation. *J. Biol. Chem.* **278,** 30091–30097 (2003).
83. Ungureanu, D. SUMO-1 conjugation selectively modulates STAT1-mediated gene responses. *Blood* **106,** 224–226 (2005).

84. Droescher, M., Begitt, A., Marg, A., Zacharias, M. & Vinkemeier, U. Cytokine-induced Paracrystals Prolong the Activity of Signal Transducers and Activators of Transcription (STAT) and Provide a Model for the Regulation of Protein Solubility by Small Ubiquitin-like Modifier (SUMO). *J. Biol. Chem.* **286,** 18731–18746 (2011).
85. Malakhov, M. P. *et al.* High-throughput Immunoblotting. UBIQUITIN-LIKE PROTEIN ISG15 MODIFIES KEY REGULATORS OF SIGNAL TRANSDUCTION. *J. Biol. Chem.* **278,** 16608–16613 (2003).
86. Malakhova, O. A. Protein ISGylation modulates the JAK-STAT signaling pathway. *Genes Dev.* **17,** 455–460 (2003).
87. Kim, K. I. *et al.* Ube1L and Protein ISGylation Are Not Essential for Alpha/Beta Interferon Signaling. *Mol. Cell. Biol.* **26,** 472–479 (2006).
88. Gewinner, C. *et al.* The coactivator of transcription CREB-binding protein interacts preferentially with the glycosylated form of Stat5. *J. Biol. Chem.* **279,** 3563–3572 (2004).
89. Wilson, K. L. & Berk, J. M. The nuclear envelope at a glance. *J. Cell Sci.* **123,** 1973–1978 (2010).
90. Schindler, C., Levy, D. E. & Decker, T. JAK-STAT Signaling: From Interferons to Cytokines. *J. Biol. Chem.* **282,** 20059–20063 (2007).
91. Stark, G. R. & Darnell, J. E. The JAK-STAT Pathway at Twenty. *Immunity* **36,** 503–514 (2012).
92. Koster, M. & Hauser, H. Dynamic redistribution of STAT1 protein in IFN signaling visualized by GFP fusion proteins. *Eur. J. Biochem.* **260,** 137–144 (1999).
93. Haspel, R. L., Salditt-Georgieff, M. & Darnell, J. E. The rapid inactivation of nuclear tyrosine phosphorylated Stat1 depends upon a protein tyrosine phosphatase. *EMBO J.* **15,** 6262–6268 (1996).
94. Ten Hoeve, J. *et al.* Identification of a Nuclear Stat1 Protein Tyrosine Phosphatase. *Mol. Cell. Biol.* **22,** 5662–5668 (2002).
95. McBride, K. M. Nuclear export signal located within theDNA-binding domain of the STAT1transcription factor. *EMBO J.* **19,** 6196–6206 (2000).
96. Sekimoto, T., Imamoto, N., Nakajima, K., Hirano, T. & Yoneda, Y. Extracellular signal-dependent nuclear import of Stat1 is mediated by nuclear pore-targeting complex formation with NPI-1, but not Rch1. *EMBO J.* **16,** 7067–7077 (1997).
97. Marg, A. *et al.* Nucleocytoplasmic shuttling by nucleoporins Nup153 and Nup214 and CRM1-dependent nuclear export control the subcellular distribution of latent Stat1. *J. Cell Biol.* **165,** 823–833 (2004).
98. Meyer, T., Begitt, A., Lödige, I., van Rossum, M. & Vinkemeier, U. Constitutive and IFN-γ-induced nuclear import of STAT1 proceed through independent pathways. *EMBO J.* **21,** 344–354 (2002).
99. Banninger, G. & Reich, N. C. STAT2 Nuclear Trafficking. *J. Biol. Chem.* **279,** 39199–39206 (2004).
100. Frahm, T., Hauser, H. & Köster, M. IFN-type-I-mediated signaling is regulated by modulation of STAT2 nuclear export. *J. Cell Sci.* **119,** 1092–1104 (2006).
101. Martinez-Moczygemba, M., Gutch, M. J., French, D. L. & Reich, N. C. Distinct STAT Structure Promotes Interaction of STAT2 with the p48 Subunit of the Interferon- -stimulated Transcription Factor ISGF3. *J. Biol. Chem.* **272,** 20070–20076 (1997).
102. Herrmann, A. *et al.* STAT3 is enriched in nuclear bodies. *J. Cell Sci.* **117,** 339–349 (2004).
103. Liu, L., McBride, K. M. & Reich, N. C. STAT3 nuclear import is independent of tyrosine phosphorylation and mediated by importin- 3. *Proc. Natl. Acad. Sci.* **102,** 8150–8155 (2005).
104. Ma, J., Zhang, T., Novotny-Diermayr, V., Tan, A. L. C. & Cao, X. A Novel Sequence in the Coiled-coil Domain of Stat3 Essential for Its Nuclear Translocation. *J. Biol. Chem.* **278,** 29252–29260 (2003).

105. Ma, J. & Cao, X. Regulation of Stat3 nuclear import by importin α5 and importin α7 via two different functional sequence elements. *Cell. Signal.* **18,** 1117–1126 (2006).
106. Cimica, V. & Reich, N. C. in *JAK-STAT Signalling* (eds. Nicholson, S. E. & Nicola, N. A.) **967,** 189–202 (Humana Press, 2013).
107. Pranada, A. L., Metz, S., Herrmann, A., Heinrich, P. C. & Muller-Newen, G. Real Time Analysis of STAT3 Nucleocytoplasmic Shuttling. *J. Biol. Chem.* **279,** 15114–15123 (2004).
108. Bhattacharya, S. & Schindler, C. Regulation of Stat3 nuclear export. *J. Clin. Invest.* **111,** 553–559 (2003).
109. Berenson, L. S., Gavrieli, M., Farrar, J. D., Murphy, T. L. & Murphy, K. M. Distinct characteristics of murine STAT4 activation in response to IL-12 and IFN-alpha. *J. Immunol. Baltim. Md 1950* **177,** 5195–5203 (2006).
110. Toyoda, H. *et al.* Impairment of IL-12-dependent STAT4 nuclear translocation in a patient with recurrent Mycobacterium avium infection. *J. Immunol. Baltim. Md 1950* **172,** 3905–3912 (2004).
111. Zeng, R., Aoki, Y., Yoshida, M., Arai, K. & Watanabe, S. Stat5B shuttles between cytoplasm and nucleus in a cytokine-dependent and -independent manner. *J. Immunol. Baltim. Md 1950* **168,** 4567–4575 (2002).
112. Iyer, J. & Reich, N. C. Constitutive nuclear import of latent and activated STAT5a by its coiled coil domain. *FASEB J.* **22,** 391–400 (2007).
113. Shin, H. Y. & Reich, N. C. Dynamic trafficking of STAT5 depends on an unconventional nuclear localization signal. *J. Cell Sci.* **126,** 3333–3343 (2013).
114. Chen, H.-C. & Reich, N. C. Live Cell Imaging Reveals Continuous STAT6 Nuclear Trafficking. *J. Immunol.* **185,** 64–70 (2010).
115. Reich, N. C. STATs get their move on. *JAKSTAT.* **2,** e27080 (2013).
116. Wenta, N., Strauss, H., Meyer, S. & Vinkemeier, U. Tyrosine phosphorylation regulates the partitioning of STAT1 between different dimer conformations. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 9238–9243 (2008).
117. Mao, X. *et al.* Structural bases of unphosphorylated STAT1 association and receptor binding. *Mol.Cell* **17,** 761–771 (2005).
118. Neculai, D. *et al.* Structure of the unphosphorylated STAT5a dimer. *J.Biol.Chem.* **280,** 40782–40787 (2005).
119. Bernado, P. *et al.* Structural characterization of unphosphorylated STAT5a oligomerization equilibrium in solution by small-angle X-ray scattering. *Protein Sci* **18,** 716–726 (2009).
120. Luker, K. E. *et al.* Kinetics of regulated protein-protein interactions revealed with firefly luciferase complementation imaging in cells and living animals. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 12288–12293 (2004).
121. Ren, Z. *et al.* Crystal structure of unphosphorylated STAT3 core fragment. *Biochem.Biophys.Res.Commun.* **374,** 1–5 (2008).
122. Vogt, M. *et al.* The role of the N-terminal domain in dimerization and nucleocytoplasmic shuttling of latent STAT3. *J. Cell Sci.* **124,** 900–909 (2011).
123. Braunstein, J., Brutsaert, S., Olson, R. & Schindler, C. STATs dimerize in the absence of phosphorylation. *J. Biol. Chem.* **278,** 34133–34140 (2003).
124. Haan, S. *et al.* Cytoplasmic STAT proteins associate prior to activation. *Biochem. J.* **345 Pt 3,** 417–421 (2000).
125. Stancato, L. F., David, M., Carter-Su, C., Larner, A. C. & Pratt, W. B. Preassociation of STAT1 with STAT2 and STAT3 in separate signalling complexes prior to cytokine stimulation. *J. Biol. Chem.* **271,** 4134–4137 (1996).
126. Shuai, K. *et al.* Interferon activation of the transcription factor Stat91 involves dimerization through SH2-phosphotyrosyl peptide interactions. *Cell* **76,** 821–828 (1994).
127. Aaronson, D. S. & Horvath, C. M. A road map for those who don't know JAK-STAT. *Science* **296,** 1653–1655 (2002).

128. Darnell, J. E. STATs and gene regulation. *Science* **277,** 1630–1635 (1997).
129. Parmar, S. & Platanias, L. C. Interferons: mechanisms of action and clinical applications. *Curr. Opin. Oncol.* **15,** 431–439 (2003).
130. Platanias, L. C. & Fish, E. N. Signaling pathways activated by interferons. *Exp. Hematol.* **27,** 1583–1592 (1999).
131. Stark, G. R., Kerr, I. M., Williams, B. R., Silverman, R. H. & Schreiber, R. D. How cells respond to interferons. *Annu. Rev. Biochem.* **67,** 227–264 (1998).
132. Becker, S., Groner, B. & Muller, C. W. Three-dimensional structure of the Stat3beta homodimer bound to DNA. *Nature* **394,** 145–151 (1998).
133. Chen, X. *et al.* Crystal structure of a tyrosine phosphorylated STAT-1 dimer bound to DNA. *Cell* **93,** 827–839 (1998).
134. Nkansah, E. *et al.* Observation of unphosphorylated STAT3 core protein binding to target dsDNA by PEMSA and X-ray crystallography. *FEBS Lett* **587,** 833–839 (2013).
135. Soler-Lopez, M. *et al.* Structure of an activated Dictyostelium STAT in its DNA-unbound form. *Mol. Cell* **13,** 791–804 (2004).
136. Meyer, T., Hendry, L., Begitt, A., John, S. & Vinkemeier, U. A single residue modulates tyrosine dephosphorylation, oligomerization, and nuclear accumulation of stat transcription factors. *J. Biol. Chem.* **279,** 18998–19007 (2004).
137. Vinkemeier, U. *et al.* DNA binding of in vitro activated Stat1 alpha, Stat1 beta and truncated Stat1: interaction between NH2-terminal domains stabilizes binding of two dimers to tandem DNA sites. *EMBO J.* **15,** 5616–5626 (1996).
138. Vinkemeier, U., Moarefi, I., Darnell, J. E., Jr. & Kuriyan, J. Structure of the amino-terminal protein interaction domain of STAT-4. *Science* **279,** 1048–1052 (1998).
139. Xu, X., Sun, Y.-L. & Hoey, T. Cooperative DNA Binding and Sequence-Selective Recognition Conferred by the STAT Amino-Terminal Domain. *Science* **273,** 794–797 (1996).
140. Meyer, W. K., Reichenbach, P., Schindler, U., Soldaini, E. & Nabholz, M. Interaction of STAT5 dimers on two low affinity binding sites mediates interleukin 2 (IL-2) stimulation of IL-2 receptor alpha gene transcription. *J. Biol. Chem.* **272,** 31821–31828 (1997).
141. Hou, Z. *et al.* Two tandemly linked interferon-gamma-activated sequence elements in the promoter of glycosylation-dependent cell adhesion molecule 1 gene synergistically respond to prolactin in mouse mammary epithelial cells. *Mol. Endocrinol. Baltim. Md* **17,** 1910–1920 (2003).
142. John, S., Vinkemeier, U., Soldaini, E., Darnell, J. E. & Leonard, W. J. The significance of tetramerization in promoter recruitment by Stat5. *Mol. Cell. Biol.* **19,** 1910–1918 (1999).
143. Moriggl, R. *et al.* Stat5 tetramer formation is associated with leukemogenesis. *Cancer Cell* **7,** 87–99 (2005).
144. Knight, R. A., Scarabelli, T. M. & Stephanou, A. STAT transcription in the ischemic heart. *JAK-STAT* **1,** 111–117 (2012).
145. Nishio, H., Matsui, K., Tsuji, H., Tamura, A. & Suzuki, K. Expression of the janus kinases-signal transducers and activators of transcription pathway in Hassall's corpuscles of the human thymus. *Histochem. Cell Biol.* **113,** 427–431 (2000).
146. Wakao, H., Gouilleux, F. & Groner, B. Mammary gland factor (MGF) is a novel member of the cytokine regulated transcription factor gene family and confers the prolactin response. *EMBO J.* **13,** 2182–2191 (1994).
147. Durbin, J. E., Hackenmiller, R., Simon, M. C. & Levy, D. E. Targeted disruption of the mouse Stat1 gene results in compromised innate immunity to viral disease. *Cell* **84,** 443–450 (1996).
148. Meraz, M. A. *et al.* Targeted disruption of the Stat1 gene in mice reveals unexpected physiologic specificity in the JAK-STAT signaling pathway. *Cell* **84,** 431–442 (1996).
149. Zhang, Q., Ekhterae, D. & Kim, K. H. Molecular cloning and characterization of P113, a mouse SNF2/SWI2-related transcription factor. *Gene* **202,** 31–37 (1997).

150. Park, C., Li, S., Cha, E. & Schindler, C. Immune response in Stat2 knockout mice. *Immunity* **13,** 795–804 (2000).
151. Zhong, Z., Wen, Z. & Darnell, J. E. Stat3 and Stat4: members of the family of signal transducers and activators of transcription. *Proc. Natl. Acad. Sci. U. S. A.* **91,** 4806–4810 (1994).
152. Takeda, K. *et al.* Targeted disruption of the mouse Stat3 gene leads to early embryonic lethality. *Proc. Natl. Acad. Sci. U. S. A.* **94,** 3801–3804 (1997).
153. Nakajima, K. *et al.* A central role for Stat3 in IL-6-induced regulation of growth and differentiation in M1 leukemia cells. *EMBO J.* **15,** 3651–3658 (1996).
154. Takeda, K. *et al.* Stat3 activation is responsible for IL-6-dependent T cell proliferation through preventing apoptosis: generation and characterization of T cell-specific Stat3-deficient mice. *J. Immunol. Baltim. Md 1950* **161,** 4652–4660 (1998).
155. Takeda, K. *et al.* Enhanced Th1 activity and development of chronic enterocolitis in mice devoid of Stat3 in macrophages and neutrophils. *Immunity* **10,** 39–49 (1999).
156. Avery, D. T. *et al.* B cell-intrinsic signaling through IL-21 receptor and STAT3 is required for establishing long-lived antibody responses in humans. *J. Exp. Med.* **207,** 155–171 (2010).
157. Zhou, L. *et al.* IL-6 programs T(H)-17 cell differentiation by promoting sequential engagement of the IL-21 and IL-23 pathways. *Nat. Immunol.* **8,** 967–974 (2007).
158. Owaki, T. *et al.* STAT3 is indispensable to IL-27-mediated cell proliferation but not to IL-27-induced Th1 differentiation and suppression of proinflammatory cytokine production. *J. Immunol. Baltim. Md 1950* **180,** 2903–2911 (2008).
159. Shimozaki, K., Nakajima, K., Hirano, T. & Nagata, S. Involvement of STAT3 in the granulocyte colony-stimulating factor-induced differentiation of myeloid cells. *J. Biol. Chem.* **272,** 25184–25189 (1997).
160. Kaplan, M. H., Sun, Y. L., Hoey, T. & Grusby, M. J. Impaired IL-12 responses and enhanced development of Th2 cells in Stat4-deficient mice. *Nature* **382,** 174–177 (1996).
161. Herrada, G. & Wolgemuth, D. J. The mouse transcription factor Stat4 is expressed in haploid male germ cells and is present in the perinuclear theca of spermatozoa. *J. Cell Sci.* **110 (Pt 14),** 1543–1553 (1997).
162. Cattaneo, E. *et al.* Activation of the JAK/STAT pathway leads to proliferation of ST14A central nervous system progenitor cells. *J. Biol. Chem.* **271,** 23374–23379 (1996).
163. Chow, J. C. *et al.* Growth hormone stimulates tyrosine phosphorylation of JAK2 and STAT5, but not insulin receptor substrate-1 or SHC proteins in liver and skeletal muscle of normal rats in vivo. *Endocrinology* **137,** 2880–2886 (1996).
164. Liu, X., Robinson, G. W., Gouilleux, F., Groner, B. & Hennighausen, L. Cloning and expression of Stat5 and an additional homologue (Stat5b) involved in prolactin signal transduction in mouse mammary tissue. *Proc. Natl. Acad. Sci. U. S. A.* **92,** 8831–8835 (1995).
165. Cui, Y. *et al.* Inactivation of Stat5 in mouse mammary epithelium during pregnancy reveals distinct functions in cell proliferation, survival, and differentiation. *Mol. Cell. Biol.* **24,** 8037–8047 (2004).
166. Teglund, S. *et al.* Stat5a and Stat5b proteins have essential and nonessential, or redundant, roles in cytokine responses. *Cell* **93,** 841–850 (1998).
167. Udy, G. B. *et al.* Requirement of STAT5b for sexual dimorphism of body growth rates and liver gene expression. *Proc. Natl. Acad. Sci. U. S. A.* **94,** 7239–7244 (1997).
168. Socolovsky, M., Fallon, A. E., Wang, S., Brugnara, C. & Lodish, H. F. Fetal anemia and apoptosis of red cell progenitors in Stat5a-/-5b-/- mice: a direct role for Stat5 in Bcl-X(L) induction. *Cell* **98,** 181–191 (1999).
169. Kieslinger, M. *et al.* Antiapoptotic activity of Stat5 required during terminal stages of myeloid differentiation. *Genes Dev.* **14,** 232–244 (2000).

170. Wang, Z., Li, G., Tse, W. & Bunting, K. D. Conditional deletion of STAT5 in adult mouse hematopoietic stem cells causes loss of quiescence and permits efficient nonablative stem cell replacement. *Blood* **113,** 4856–4865 (2009).
171. Yao, Z. *et al.* Stat5a/b are essential for normal lymphoid development and differentiation. *Proc. Natl. Acad. Sci.* **103,** 1000–1005 (2006).
172. Shimoda, K. *et al.* Lack of IL-4-induced Th2 response and IgE class switching in mice with disrupted Stat6 gene. *Nature* **380,** 630–633 (1996).
173. Takeda, K. *et al.* Essential role of Stat6 in IL-4 signalling. *Nature* **380,** 627–630 (1996).
174. Urban, J. F. *et al.* IL-13, IL-4Ralpha, and Stat6 are required for the expulsion of the gastrointestinal nematode parasite Nippostrongylus brasiliensis. *Immunity* **8,** 255–264 (1998).
175. Cheon, H., Yang, J. & Stark, G. R. The functions of signal transducers and activators of transcriptions 1 and 3 as cytokine-inducible proteins. *J. Interferon Cytokine Res. Off. J. Int. Soc. Interferon Cytokine Res.* **31,** 33–40 (2011).
176. Yue, H., Li, W., Desnoyer, R. & Karnik, S. S. Role of nuclear unphosphorylated STAT3 in angiotensin II type 1 receptor-induced cardiac hypertrophy. *Cardiovasc. Res.* **85,** 90–99 (2010).
177. Testoni, B. *et al.* Chromatin Dynamics of Gene Activation and Repression in Response to Interferon (IFN) Reveal New Roles for Phosphorylated and Unphosphorylated Forms of the Transcription Factor STAT2. *J. Biol. Chem.* **286,** 20217–20227 (2011).
178. McGuckin, C. P. *et al.* Ischemic brain injury: A consortium analysis of key factors involved in mesenchymal stem cell-mediated inflammatory reduction. *Arch. Biochem. Biophys.* **534,** 88–97 (2013).
179. Christova, R. *et al.* P-STAT1 mediates higher-order chromatin remodelling of the human MHC in response to IFNgamma. *J. Cell Sci.* **120,** 3262–3270 (2007).
180. Gough, D. J. *et al.* Mitochondrial STAT3 supports Ras-dependent oncogenic transformation. *Science* **324,** 1713–1716 (2009).
181. Meier, J. A. & Larner, A. C. Toward a new STATe: The role of STATs in mitochondrial function. *Semin. Immunol.* **26,** 20–28 (2014).
182. Wegrzyn, J. *et al.* Function of Mitochondrial Stat3 in Cellular Respiration. *Science* **323,** 793–797 (2009).
183. Lee, J. E. *et al.* Nongenomic STAT5-dependent effects on Golgi apparatus and endoplasmic reticulum structure and function. *Am. J. Physiol. Cell Physiol.* **302,** C804–820 (2012).
184. Bradley, H. L., Hawley, T. S. & Bunting, K. D. Cell intrinsic defects in cytokine responsiveness of STAT5-deficient hematopoietic stem cells. *Blood* **100,** 3983–3989 (2002).
185. Bradley, H. L., Couldrey, C. & Bunting, K. D. Hematopoietic-repopulating defects from STAT5-deficient bone marrow are not fully accounted for by loss of thrombopoietin responsiveness. *Blood* **103,** 2965–2972 (2004).
186. Bunting, K. D. *et al.* Reduced lymphomyeloid repopulating activity from adult bone marrow and fetal liver of mice lacking expression of STAT5. *Blood* **99,** 479–487 (2002).
187. Couldrey, C., Bradley, H. L. & Bunting, K. D. A STAT5 modifier locus on murine chromosome 7 modulates engraftment of hematopoietic stem cells during steady-state hematopoiesis. *Blood* **105,** 1476–1483 (2005).
188. Grimley, P. M., Dong, F. & Rui, H. Stat5a and Stat5b: fraternal twins of signal transduction and transcriptional activation. *Cytokine Growth Factor Rev* **10,** 131–157 (1999).
189. Li, G. *et al.* STAT5 requires the N-domain to maintain hematopoietic stem cell repopulating function and appropriate lymphoid-myeloid lineage output. *Exp. Hematol.* **35,** 1684–1694 (2007).
190. Dai, X. *et al.* Stat5 is essential for early B cell development but not for B cell maturation and function. *J. Immunol. Baltim. Md 1950* **179,** 1068–1079 (2007).

191. Li, G. *et al.* Gab2 promotes hematopoietic stem cell maintenance and self-renewal synergistically with STAT5. *PloS One* **5,** e9152 (2010).
192. Snow, J. W. *et al.* STAT5 promotes multilineage hematolymphoid development in vivo through effects on early hematopoietic progenitor cells. *Blood* **99,** 95–101 (2002).
193. Barnstein, B. O. *et al.* Stat5 expression is required for IgE-mediated mast cell function. *J. Immunol. Baltim. Md 1950* **177,** 3421–3426 (2006).
194. Moriggl, R. *et al.* Stat5 Is Required for IL-2-Induced Cell Cycle Progression of Peripheral T Cells. *Immunity* **10,** 249–259 (1999).
195. Ohmori, K. *et al.* IL-3 induces basophil expansion in vivo by directing granulocyte-monocyte progenitors to differentiate into basophil lineage-restricted progenitors in the bone marrow and by increasing the number of basophil/mast cell progenitors in the spleen. *J. Immunol. Baltim. Md 1950* **182,** 2835–2841 (2009).
196. Shelburne, C. P. *et al.* Stat5 expression is critical for mast cell development and survival. *Blood* **102,** 1290–1297 (2003).
197. Moriggl, R., Sexl, V., Piekorz, R., Topham, D. & Ihle, J. N. Stat5 Activation Is Uniquely Associated with Cytokine Signaling in Peripheral T Cells. *Immunity* **11,** 225–230 (1999).
198. Peschon, J. J. *et al.* Early lymphocyte expansion is severely impaired in interleukin 7 receptor-deficient mice. *J. Exp. Med.* **180,** 1955–1960 (1994).
199. Frasor, J. *et al.* Differential roles for signal transducers and activators of transcription 5a and 5b in PRL stimulation of ERalpha and ERbeta transcription. *Mol. Endocrinol. Baltim. Md* **15,** 2172–2181 (2001).
200. Bernasconi, A. *et al.* Characterization of immunodeficiency in a patient with growth hormone insensitivity secondary to a novel STAT5b gene mutation. *Pediatrics* **118,** e1584–1592 (2006).
201. Cohen, A. C. *et al.* Cutting edge: Decreased accumulation and regulatory function of CD4+ CD25(high) T cells in human STAT5b deficiency. *J. Immunol. Baltim. Md 1950* **177,** 2770–2774 (2006).
202. Vidarsdottir, S. *et al.* Clinical and biochemical characteristics of a male patient with a novel homozygous STAT5b mutation. *J. Clin. Endocrinol. Metab.* **91,** 3482–3485 (2006).
203. Burchill, M. A., Yang, J., Vogtenhuber, C., Blazar, B. R. & Farrar, M. A. IL-2 receptor beta-dependent STAT5 activation is required for the development of Foxp3+ regulatory T cells. *J. Immunol. Baltim. Md 1950* **178,** 280–290 (2007).
204. Jenks, J. A. *et al.* Differentiating the roles of STAT5B and STAT5A in human CD4+ T cells. *Clin. Immunol. Orlando Fla* **148,** 227–236 (2013).
205. Yao, Z. *et al.* Nonredundant roles for Stat5a/b in directly regulating Foxp3. *Blood* **109,** 4368–4375 (2007).
206. Dumon, S. *et al.* IL-3 dependent regulation of Bcl-xL gene expression by STAT5 in a bone marrow derived cell line. *Oncogene* **18,** 4191–4199 (1999).
207. Nosaka, T. *et al.* STAT5 as a molecular regulator of proliferation, differentiation and apoptosis in hematopoietic cells. *EMBO J.* **18,** 4754–4765 (1999).
208. Bacon, C. M. *et al.* Thrombopoietin (TPO) induces tyrosine phosphorylation and activation of STAT5 and STAT3. *FEBS Lett.* **370,** 63–68 (1995).
209. Drachman, J. G., Sabath, D. F., Fox, N. E. & Kaushansky, K. Thrombopoietin signal transduction in purified murine megakaryocytes. *Blood* **89,** 483–492 (1997).
210. Kirito, K. *et al.* Thrombopoietin regulates Bcl-xL gene expression through Stat5 and phosphatidylinositol 3-kinase activation pathways. *J. Biol. Chem.* **277,** 8329–8337 (2002).
211. Ooi, J., Tojo, A., Asano, S., Sato, Y. & Oka, Y. Thrombopoietin induces tyrosine phosphorylation of a common beta subunit of GM-CSF receptor and its association with Stat5 in TF-1/TPO cells. *Biochem. Biophys. Res. Commun.* **246,** 132–136 (1998).

212. Schulze, H., Ballmaier, M., Welte, K. & Germeshausen, M. Thrombopoietin induces the generation of distinct Stat1, Stat3, Stat5a and Stat5b homo- and heterodimeric complexes with different kinetics in human platelets. *Exp. Hematol.* **28,** 294–304 (2000).
213. Han, L. *et al.* Single-cell STAT5 signal transduction profiling in normal and leukemic stem and progenitor cell populations reveals highly distinct cytokine responses. *PloS One* **4,** e7989 (2009).
214. Liu, F. *et al.* Csf3r mutations in mice confer a strong clonal HSC advantage via activation of Stat5. *J. Clin. Invest.* **118,** 946–955 (2008).
215. Kimura, A. *et al.* The transcription factors STAT5A/B regulate GM-CSF-mediated granulopoiesis. *Blood* **114,** 4721–4728 (2009).
216. Ko, J. S. *et al.* Direct and differential suppression of myeloid-derived suppressor cell subsets by sunitinib is compartmentally constrained. *Cancer Res.* **70,** 3526–3536 (2010).
217. Mui, A. L., Wakao, H., O'Farrell, A. M., Harada, N. & Miyajima, A. Interleukin-3, granulocyte-macrophage colony stimulating factor and interleukin-5 transduce signals through two STAT5 homologs. *EMBO J.* **14,** 1166–1175 (1995).
218. Qi, X. *et al.* Antagonistic regulation by the transcription factors C/EBPα and MITF specifies basophil and mast cell fates. *Immunity* **39,** 97–110 (2013).
219. Oda, A. *et al.* Erythropoietin induces tyrosine phosphorylation of Jak2, STAT5A, and STAT5B in primary cultured human erythroid precursors. *Blood* **92,** 443–451 (1998).
220. Silva, M. *et al.* Erythropoietin can induce the expression of bcl-x(L) through Stat5 in erythropoietin-dependent progenitor cell lines. *J. Biol. Chem.* **274,** 22165–22169 (1999).
221. Boer, A.-K., Drayer, A. L. & Vellenga, E. Stem cell factor enhances erythropoietin-mediated transactivation of signal transducer and activator of transcription 5 (STAT5) via the PKA/CREB pathway. *Exp. Hematol.* **31,** 512–520 (2003).
222. Hoelbl, A. *et al.* Clarifying the role of Stat5 in lymphoid development and Abelson-induced transformation. *Blood* **107,** 4898–4906 (2006).
223. Kondo, M., Akashi, K., Domen, J., Sugamura, K. & Weissman, I. L. Bcl-2 rescues T lymphopoiesis, but not B or NK cell development, in common gamma chain-deficient mice. *Immunity* **7,** 155–162 (1997).
224. Maraskovsky, E. *et al.* Bcl-2 can rescue T lymphocyte development in interleukin-7 receptor-deficient mice but not in mutant rag-1-/- mice. *Cell* **89,** 1011–1019 (1997).
225. Maraskovsky, E., Peschon, J. J., McKenna, H., Teepe, M. & Strasser, A. Overexpression of Bcl-2 does not rescue impaired B lymphopoiesis in IL-7 receptor-deficient mice but can enhance survival of mature B cells. *Int. Immunol.* **10,** 1367–1375 (1998).
226. Lin, H. & Grosschedl, R. Failure of B-cell differentiation in mice lacking the transcription factor EBF. *Nature* **376,** 263–267 (1995).
227. Urbánek, P., Wang, Z. Q., Fetka, I., Wagner, E. F. & Busslinger, M. Complete block of early B cell differentiation and altered patterning of the posterior midbrain in mice lacking Pax5/BSAP. *Cell* **79,** 901–912 (1994).
228. Kikuchi, K., Lai, A. Y., Hsu, C.-L. & Kondo, M. IL-7 receptor signaling is necessary for stage transition in adult B cell development through up-regulation of EBF. *J. Exp. Med.* **201,** 1197–1203 (2005).
229. Roessler, S. *et al.* Distinct promoters mediate the regulation of Ebf1 gene expression by interleukin-7 and Pax5. *Mol. Cell. Biol.* **27,** 579–594 (2007).
230. Goetz, C. A. *et al.* Restricted STAT5 activation dictates appropriate thymic B versus T cell lineage commitment. *J. Immunol. Baltim. Md 1950* **174,** 7753–7763 (2005).
231. Hirokawa, S., Sato, H., Kato, I. & Kudo, A. EBF-regulating Pax5 transcription is enhanced by STAT5 in the early stage of B cells. *Eur. J. Immunol.* **33,** 1824–1829 (2003).
232. Decker, T. *et al.* Stepwise activation of enhancer and promoter regions of the B cell commitment gene Pax5 in early lymphopoiesis. *Immunity* **30,** 508–520 (2009).

233. Scheeren, F. A. *et al.* STAT5 regulates the self-renewal capacity and differentiation of human memory B cells and controls Bcl-6 expression. *Nat. Immunol.* **6,** 303–313 (2005).
234. Bertolino, E. *et al.* Regulation of interleukin 7-dependent immunoglobulin heavy-chain variable gene rearrangements by transcription factor STAT5. *Nat. Immunol.* **6,** 836–843 (2005).
235. Hewitt, S. L. *et al.* RAG-1 and ATM coordinate monoallelic recombination and nuclear positioning of immunoglobulin loci. *Nat. Immunol.* **10,** 655–664 (2009).
236. Malin, S. *et al.* Role of STAT5 in controlling cell survival and immunoglobulin gene recombination during pro-B cell development. *Nat. Immunol.* **11,** 171–179 (2010).
237. Johnson, K. *et al.* Regulation of immunoglobulin light-chain recombination by the transcription factor IRF-4 and the attenuation of interleukin-7 signaling. *Immunity* **28,** 335–345 (2008).
238. Malin, S., McManus, S. & Busslinger, M. STAT5 in B cell development and leukemia. *Curr. Opin. Immunol.* **22,** 168–176 (2010).
239. Burchill, M. A. *et al.* Distinct effects of STAT5 activation on CD4+ and CD8+ T cell homeostasis: development of CD4+CD25+ regulatory T cells versus CD8+ memory T cells. *J. Immunol. Baltim. Md 1950* **171,** 5853–5864 (2003).
240. Park, J.-H. *et al.* 'Coreceptor tuning': cytokine signals transcriptionally tailor CD8 coreceptor expression to the self-specificity of the TCR. *Nat. Immunol.* **8,** 1049–1059 (2007).
241. Park, J.-H. *et al.* Signaling by intrathymic cytokines, not T cell antigen receptors, specifies CD8 lineage choice and promotes the differentiation of cytotoxic-lineage T cells. *Nat. Immunol.* **11,** 257–264 (2010).
242. Mitchell, D. M. & Williams, M. A. Disparate Roles for STAT5 in Primary and Secondary CTL Responses. *J. Immunol.* **190,** 3390–3398 (2013).
243. Vogtenhuber, C. *et al.* Constitutively active Stat5b in CD4+ T cells inhibits graft-versus-host disease lethality associated with increased regulatory T-cell potency and decreased T effector cell responses. *Blood* **116,** 466–474 (2010).
244. Zhu, J., Cote-Sierra, J., Guo, L. & Paul, W. E. Stat5 activation plays a critical role in Th2 differentiation. *Immunity* **19,** 739–748 (2003).
245. Laurence, A. *et al.* Interleukin-2 signaling via STAT5 constrains T helper 17 cell generation. *Immunity* **26,** 371–381 (2007).
246. Lenardo, M. J. Interleukin-2 programs mouse alpha beta T lymphocytes for apoptosis. *Nature* **353,** 858–861 (1991).
247. Sadlack, B. *et al.* Ulcerative colitis-like disease in mice with a disrupted interleukin-2 gene. *Cell* **75,** 253–261 (1993).
248. Suzuki, H. *et al.* Deregulated T cell activation and autoimmunity in mice lacking interleukin-2 receptor beta. *Science* **268,** 1472–1476 (1995).
249. Willerford, D. M. *et al.* Interleukin-2 receptor alpha chain regulates the size and content of the peripheral lymphoid compartment. *Immunity* **3,** 521–530 (1995).
250. Antov, A., Yang, L., Vig, M., Baltimore, D. & Van Parijs, L. Essential role for STAT5 signaling in CD25+CD4+ regulatory T cell homeostasis and the maintenance of self-tolerance. *J. Immunol. Baltim. Md 1950* **171,** 3435–3441 (2003).
251. Snow, J. W. *et al.* Loss of tolerance and autoimmunity affecting multiple organs in STAT5A/5B-deficient mice. *J. Immunol. Baltim. Md 1950* **171,** 5042–5050 (2003).
252. Burchill, M. A. *et al.* Linked T cell receptor and cytokine signaling govern the development of the regulatory T cell repertoire. *Immunity* **28,** 112–121 (2008).
253. Lio, C.-W. J. & Hsieh, C.-S. A two-step process for thymic regulatory T cell development. *Immunity* **28,** 100–111 (2008).
254. Moran, A. E. *et al.* T cell receptor signal strength in Treg and iNKT cell development demonstrated by a novel fluorescent reporter mouse. *J. Exp. Med.* **208,** 1279–1289 (2011).

255. Vang, K. B. *et al.* IL-2, -7, and -15, but not thymic stromal lymphopoeitin, redundantly govern CD4+Foxp3+ regulatory T cell development. *J. Immunol. Baltim. Md 1950* **181,** 3285–3290 (2008).
256. Lio, C.-W. J., Dodson, L. F., Deppong, C. M., Hsieh, C.-S. & Green, J. M. CD28 facilitates the generation of Foxp3(-) cytokine responsive regulatory T cell precursors. *J. Immunol. Baltim. Md 1950* **184,** 6007–6013 (2010).
257. Vang, K. B. *et al.* Cutting edge: CD28 and c-Rel-dependent pathways initiate regulatory T cell development. *J. Immunol. Baltim. Md 1950* **184,** 4074–4077 (2010).
258. Mahmud, S. A., Manlove, L. S. & Farrar, M. A. Interleukin-2 and STAT5 in regulatory T cell development and function. *JAK-STAT* **2,** e23154 (2013).
259. Shan, J. *et al.* Interplay between mTOR and STAT5 signaling modulates the balance between regulatory and effective T cells. *Immunobiology* **220,** 510–517 (2015).
260. Chueh, F.-Y., Leong, K.-F. & Yu, C.-L. Mitochondrial translocation of signal transducer and activator of transcription 5 (STAT5) in leukemic T cells and cytokine-stimulated cells. *Biochem. Biophys. Res. Commun.* **402,** 778–783 (2010).
261. Fatrai, S., Wierenga, A. T. J., Daenen, S. M. G. J., Vellenga, E. & Schuringa, J. J. Identification of HIF2 as an important STAT5 target gene in human hematopoietic stem cells. *Blood* **117,** 3320–3330 (2011).
262. Wofford, J. A., Wieman, H. L., Jacobs, S. R., Zhao, Y. & Rathmell, J. C. IL-7 promotes Glut1 trafficking and glucose uptake via STAT5-mediated activation of Akt to support T-cell survival. *Blood* **111,** 2101–2111 (2008).
263. Niwa, H., Burdon, T., Chambers, I. & Smith, A. Self-renewal of pluripotent embryonic stem cells is mediated via activation of STAT3. *Genes Dev.* **12,** 2048–2060 (1998).
264. Hughes, K. & Watson, C. J. The spectrum of STAT functions in mammary gland development. *JAK-STAT* **1,** 151–158 (2012).
265. Watson, C. J., Gordon, K. E., Robertson, M. & Clark, A. J. Interaction of DNA-binding proteins with a milk protein gene promoter in vitro: identification of a mammary gland-specific factor. *Nucleic Acids Res.* **19,** 6603–6610 (1991).
266. Burdon, T. G., Maitland, K. A., Clark, A. J., Wallace, R. & Watson, C. J. Regulation of the sheep beta-lactoglobulin gene by lactogenic hormones is mediated by a transcription factor that binds an interferon-gamma activation site-related element. *Mol. Endocrinol. Baltim. Md* **8,** 1528–1536 (1994).
267. Li, S. & Rosen, J. M. Nuclear factor I and mammary gland factor (STAT5) play a critical role in regulating rat whey acidic protein gene expression in transgenic mice. *Mol. Cell. Biol.* **15,** 2063–2070 (1995).
268. Happ, B. & Groner, B. The activated mammary gland specific nuclear factor (MGF) enhances in vitro transcription of the β-casein gene promoter. *J. Steroid Biochem. Mol. Biol.* **47,** 21–30 (1993).
269. Ormandy, C. J. *et al.* Investigation of the transcriptional changes underlying functional defects in the mammary glands of prolactin receptor knockout mice. *Recent Prog. Horm. Res.* **58,** 297–323 (2003).
270. Liu, X. *et al.* Stat5a is mandatory for adult mammary gland development and lactogenesis. *Genes Dev.* **11,** 179–186 (1997).
271. Yamaji, D. *et al.* Development of mammary luminal progenitor cells is controlled by the transcription factor STAT5A. *Genes Dev.* **23,** 2382–2387 (2009).
272. Choi, Y. S., Chakrabarti, R., Escamilla-Hernandez, R. & Sinha, S. Elf5 conditional knockout mice reveal its role as a master regulator in mammary alveolar development: Failure of Stat5 activation and functional differentiation in the absence of Elf5. *Dev. Biol.* **329,** 227–241 (2009).
273. Arkun, Y. & Gur, M. Combining Optimal Control Theory and Molecular Dynamics for Protein Folding. *PLoS ONE* **7,** e29628 (2012).

274. Sanz-Moreno, A. *et al.* Miz1 Deficiency in the Mammary Gland Causes a Lactation Defect by Attenuated Stat5 Expression and Phosphorylation. *PLoS ONE* **9,** e89187 (2014).
275. Schmidt, J. W. *et al.* Stat5 Regulates the Phosphatidylinositol 3-Kinase/Akt1 Pathway during Mammary Gland Development and Tumorigenesis. *Mol. Cell. Biol.* **34,** 1363–1377 (2014).
276. Gao, Q. *et al.* Disruption of neural signal transducer and activator of transcription 3 causes obesity, diabetes, infertility, and thermal dysregulation. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 4661–4666 (2004).
277. Lee, J.-Y. *et al.* Loss of cytokine-STAT5 signaling in the CNS and pituitary gland alters energy balance and leads to obesity. *PloS One* **3,** e1639 (2008).
278. Schwartz, M. W., Woods, S. C., Porte, D., Seeley, R. J. & Baskin, D. G. Central nervous system control of food intake. *Nature* **404,** 661–671 (2000).
279. Villanueva, E. C. & Myers, M. G. Leptin receptor signaling and the regulation of mammalian physiology. *Int. J. Obes. 2005* **32 Suppl 7,** S8–12 (2008).
280. Wada, N. *et al.* Leptin and its receptors. *J. Chem. Neuroanat.* **61-62,** 191–199 (2014).
281. Magni, P. *et al.* Leukemia inhibitory factor induces the chemomigration of immortalized gonadotropin-releasing hormone neurons through the independent activation of the Janus kinase/signal transducer and activator of transcription 3, mitogen-activated protein kinase/extracellularly regulated kinase 1/2, and phosphatidylinositol 3-kinase/Akt signaling pathways. *Mol. Endocrinol. Baltim. Md* **21,** 1163–1174 (2007).
282. Wu, S. *et al.* Jak2 Is Necessary for Neuroendocrine Control of Female Reproduction. *J. Neurosci.* **31,** 184–192 (2011).
283. Anderson, G. M. *et al.* Suppression of Prolactin-Induced Signal Transducer and Activator of Transcription 5b Signaling and Induction of Suppressors of Cytokine Signaling Messenger Ribonucleic Acid in the Hypothalamic Arcuate Nucleus of the Rat during Late Pregnancy and Lactation. *Endocrinology* **147,** 4996–5005 (2006).
284. Buntin, J. D. & Buntin, L. Increased STAT5 signaling in the ring dove brain in response to prolactin administration and spontaneous elevations in prolactin during the breeding cycle. *Gen. Comp. Endocrinol.* **200,** 1–9 (2014).
285. Ma, F. Y. *et al.* Prolactin Specifically Activates Signal Transducer and Activator of Transcription 5b in Neuroendocrine Dopaminergic Neurons. *Endocrinology* **146,** 5112–5119 (2005).
286. Yip, S. H., Eguchi, R., Grattan, D. R. & Bunn, S. J. Prolactin signalling in the mouse hypothalamus is primarily mediated by signal transducer and activator of transcription factor 5b but not 5a. *J. Neuroendocrinol.* **24,** 1484–1491 (2012).
287. Stewart, W. C., Pearcy, L. A., Floyd, Z. E. & Stephens, J. M. STAT5A expression in Swiss 3T3 cells promotes adipogenesis in vivo in an athymic mice model system. *Obes. Silver Spring Md* **19,** 1731–1734 (2011).
288. Stewart, W. C., Morrison, R. F., Young, S. L. & Stephens, J. M. Regulation of signal transducers and activators of transcription (STATs) by effectors of adipogenesis: coordinate regulation of STATs 1, 5A, and 5B with peroxisome proliferator-activated receptor-gamma and C/AAAT enhancer binding protein-alpha. *Biochim. Biophys. Acta* **1452,** 188–196 (1999).
289. Floyd, Z. E. & Stephens, J. M. STAT5A promotes adipogenesis in nonprecursor cells and associates with the glucocorticoid receptor during adipocyte differentiation. *Diabetes* **52,** 308–314 (2003).
290. Stewart, W. C., Baugh, J. E., Floyd, Z. E. & Stephens, J. M. STAT 5 activators can replace the requirement of FBS in the adipogenesis of 3T3-L1 cells. *Biochem. Biophys. Res. Commun.* **324,** 355–359 (2004).
291. Kawai, M. *et al.* Growth hormone stimulates adipogenesis of 3T3-L1 cells through activation of the Stat5A/5B-PPARgamma pathway. *J. Mol. Endocrinol.* **38,** 19–34 (2007).

292. Shang, C. A. & Waters, M. J. Constitutively active signal transducer and activator of transcription 5 can replace the requirement for growth hormone in adipogenesis of 3T3-F442A preadipocytes. *Mol. Endocrinol. Baltim. Md* **17,** 2494–2508 (2003).
293. Yarwood, S. J. *et al.* Growth hormone-dependent differentiation of 3T3-F442A preadipocytes requires Janus kinase/signal transducer and activator of transcription but not mitogen-activated protein kinase or p70 S6 kinase signaling. *J. Biol. Chem.* **274,** 8662–8668 (1999).
294. Meirhaeghe, A. *et al.* A functional polymorphism in a STAT5B site of the human PPAR gamma 3 gene promoter affects height and lipid metabolism in a French population. *Arterioscler. Thromb. Vasc. Biol.* **23,** 289–294 (2003).
295. Wakao, H., Wakao, R., Oda, A. & Fujita, H. Constitutively active Stat5A and Stat5B promote adipogenesis. *Environ. Health Prev. Med.* **16,** 247–252 (2011).
296. Baugh, J. E., Floyd, Z. E. & Stephens, J. M. The modulation of STAT5A/GR complexes during fat cell differentiation and in mature adipocytes. *Obes. Silver Spring Md* **15,** 583–590 (2007).
297. Nanbu-Wakao, R. *et al.* Stimulation of 3T3-L1 adipogenesis by signal transducer and activator of transcription 5. *Mol. Endocrinol. Baltim. Md* **16,** 1565–1576 (2002).
298. Siersbæk, R. *et al.* Extensive chromatin remodelling and establishment of transcription factor 'hotspots' during early adipogenesis. *EMBO J.* **30,** 1459–1472 (2011).
299. Kofoed, E. M. *et al.* Growth hormone insensitivity associated with a STAT5b mutation. *N. Engl. J. Med.* **349,** 1139–1147 (2003).
300. Chia, D. J. *et al.* Characterization of distinct Stat5b binding sites that mediate growth hormone-stimulated IGF-I gene transcription. *J. Biol. Chem.* **281,** 3190–3197 (2006).
301. Chia, D. J., Varco-Merth, B. & Rotwein, P. Dispersed Chromosomal Stat5b-binding elements mediate growth hormone-activated insulin-like growth factor-I gene transcription. *J. Biol. Chem.* **285,** 17636–17647 (2010).
302. Eleswarapu, S., Ge, X., Wang, Y., Yu, J. & Jiang, H. Growth hormone-activated STAT5 may indirectly stimulate IGF-I gene transcription through HNF-3{gamma}. *Mol. Endocrinol. Baltim. Md* **23,** 2026–2037 (2009).
303. Hosui, A. *et al.* Loss of STAT5 causes liver fibrosis and cancer development through increased TGF-{beta} and STAT3 activation. *J. Exp. Med.* **206,** 819–831 (2009).
304. Yu, J. H. *et al.* The transcription factors signal transducer and activator of transcription 5A (STAT5A) and STAT5B negatively regulate cell proliferation through the activation of cyclin-dependent kinase inhibitor 2b (Cdkn2b) and Cdkn1a expression. *Hepatol. Baltim. Md* **52,** 1808–1818 (2010).
305. Calabrese, V. *et al.* SOCS1 links cytokine signaling to p53 and senescence. *Mol. Cell* **36,** 754–767 (2009).
306. Mallette, F. A., Gaumont-Leclerc, M.-F. & Ferbeyre, G. The DNA damage signaling pathway is a critical mediator of oncogene-induced senescence. *Genes Dev.* **21,** 43–48 (2007).
307. Waxman, D. J. & O'Connor, C. Growth hormone regulation of sex-dependent liver gene expression. *Mol. Endocrinol. Baltim. Md* **20,** 2613–2629 (2006).
308. Blaas, L. *et al.* Disruption of the growth hormone--signal transducer and activator of transcription 5--insulinlike growth factor 1 axis severely aggravates liver fibrosis in a mouse model of cholestasis. *Hepatol. Baltim. Md* **51,** 1319–1326 (2010).
309. Clodfelter, K. H. *et al.* Sex-dependent liver gene expression is extensive and largely dependent upon signal transducer and activator of transcription 5b (STAT5b): STAT5b-dependent activation of male genes and repression of female genes revealed by microarray analysis. *Mol. Endocrinol. Baltim. Md* **20,** 1333–1351 (2006).
310. Holloway, M. G. *et al.* Loss of sexually dimorphic liver gene expression upon hepatocyte-specific deletion of Stat5a-Stat5b locus. *Endocrinology* **148,** 1977–1986 (2007).

311. Van de Steeg, E. *et al.* Organic anion transporting polypeptide 1a/1b-knockout mice provide insights into hepatic handling of bilirubin, bile acids, and drugs. *J. Clin. Invest.* **120,** 2942–2952 (2010).
312. Abbaszade, I. G., Clarke, T. R., Park, C. H. & Payne, A. H. The mouse 3 beta-hydroxysteroid dehydrogenase multigene family includes two functionally distinct groups of proteins. *Mol. Endocrinol. Baltim. Md* **9,** 1214–1222 (1995).
313. Schwarz, M., Lund, E. G. & Russell, D. W. Two 7 alpha-hydroxylase enzymes in bile acid biosynthesis. *Curr. Opin. Lipidol.* **9,** 113–118 (1998).
314. Clodfelter, K. H. *et al.* Role of STAT5a in regulation of sex-specific gene expression in female but not male mouse liver revealed by microarray analysis. *Physiol. Genomics* **31,** 63–74 (2007).
315. Falany, J. L., Greer, H., Kovacs, T., Sorscher, E. J. & Falany, C. N. Elevation of hepatic sulphotransferase activities in mice with resistance to cystic fibrosis. *Biochem. J.* **364,** 115–120 (2002).
316. Nowell, P. C. & Hungerford, D. A. A minute chromosome in human chronic granulocytic leukemia. *Science* **132,** 1488–1501 (1960).
317. Rowley, J. D. Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature* **243,** 290–293 (1973).
318. Campo, E. *et al.* The 2008 WHO classification of lymphoid neoplasms and beyond: evolving concepts and practical applications. *Blood* **117,** 5019–5032 (2011).
319. Breccia, M. & Alimena, G. How to treat CML patients in the tyrosine kinase inhibitors era? From imatinib standard dose to second generation drugs front-line: unmet needs, pitfalls and advantages. *Cancer Lett.* **322,** 127–132 (2012).
320. Deininger, M. W., Goldman, J. M., Lydon, N. & Melo, J. V. The tyrosine kinase inhibitor CGP57148B selectively inhibits the growth of BCR-ABL-positive cells. *Blood* **90,** 3691–3698 (1997).
321. Druker, B. J. *et al.* Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *N. Engl. J. Med.* **344,** 1031–1037 (2001).
322. Jabbour, E. & Kantarjian, H. Chronic myeloid leukemia: 2012 update on diagnosis, monitoring, and management. *Am. J. Hematol.* **87,** 1037–1045 (2012).
323. Lindauer, M. & Hochhaus, A. Dasatinib. *Recent Results Cancer Res. Fortschritte Krebsforsch. Prog. Dans Rech. Sur Cancer* **184,** 83–102 (2010).
324. Kantarjian, H. *et al.* Nilotinib in imatinib-resistant CML and Philadelphia chromosome-positive ALL. *N. Engl. J. Med.* **354,** 2542–2551 (2006).
325. Puttini, M. *et al.* In vitro and In vivo Activity of SKI-606, a Novel Src-Abl Inhibitor, against Imatinib-Resistant Bcr-Abl+ Neoplastic Cells. *Cancer Res.* **66,** 11314–11322 (2006).
326. Frankfurt, O. & Licht, J. D. Ponatinib--A Step Forward in Overcoming Resistance in Chronic Myeloid Leukemia. *Clin. Cancer Res.* **19,** 5828–5834 (2013).
327. Cortes, J., Goldman, J. M. & Hughes, T. Current issues in chronic myeloid leukemia: monitoring, resistance, and functional cure. *J. Natl. Compr. Cancer Netw. JNCCN* **10 Suppl 3,** S1–S13 (2012).
328. Warsch, W. *et al.* High STAT5 levels mediate imatinib resistance and indicate disease progression in chronic myeloid leukemia. *Blood* **117,** 3409–3420 (2011).
329. Valent, P. Emerging stem cell concepts for imatinib-resistant chronic myeloid leukaemia: implications for the biology, management, and therapy of the disease. *Br. J. Haematol.* **142,** 361–378 (2008).
330. Valent, P. *et al.* Cancer stem cell definitions and terminology: the devil is in the details. *Nat. Rev. Cancer* **12,** 767–775 (2012).
331. Warsch, W., Grundschober, E. & Sexl, V. Adding a new facet to STAT5 in CML: multitasking for leukemic cells. *Cell Cycle Georget. Tex* **12,** 1813–1814 (2013).

332. Hoelbl, A. *et al.* Stat5 is indispensable for the maintenance of *bcr/abl* -positive leukaemia: Stat5 in leukaemia maintenance. *EMBO Mol. Med.* **2,** 98–110 (2010).
333. Kaymaz, B. T. *et al.* Suppression of STAT5A and STAT5B chronic myeloid leukemia cells via siRNA and antisense-oligonucleotide applications with the induction of apoptosis. *Am. J. Blood Res.* **3,** 58–70 (2013).
334. Kosova, B. *et al.* Suppression of STAT5A increases chemotherapeutic sensitivity in imatinib-resistant and imatinib-sensitive K562 cells. *Leuk. Lymphoma* **51,** 1895–1901 (2010).
335. Chatain, N. *et al.* Src family kinases mediate cytoplasmic retention of activated STAT5 in BCR–ABL-positive cells. *Oncogene* **32,** 3587–3597 (2013).
336. Harir, N. *et al.* Constitutive activation of Stat5 promotes its cytoplasmic localization and association with PI3-kinase in myeloid leukemias. *Blood* **109,** 1678–1686 (2007).
337. Warsch, W. *et al.* STAT5 triggers BCR-ABL1 mutation by mediating ROS production in chronic myeloid leukaemia. *Oncotarget* **3,** 1669–1687 (2013).
338. Walz, C. *et al.* Essential role for Stat5a/b in myeloproliferative neoplasms induced by BCR-ABL1 and JAK2V617F in mice. *Blood* **119,** 3550–3560 (2012).
339. Casetti, L. *et al.* Differential contributions of STAT5A and STAT5B to stress protection and tyrosine kinase inhibitor resistance of chronic myeloid leukemia stem/progenitor cells. *Cancer Res* (2013). doi:10.1158/0008-5472.CAN-12-3955
340. Galli, S. J., Tsai, M., Wershil, B. K., Tam, S. Y. & Costa, J. J. Regulation of mouse and human mast cell development, survival and function by stem cell factor, the ligand for the c-kit receptor. *Int. Arch. Allergy Immunol.* **107,** 51–53 (1995).
341. Blume-Jensen, P. *et al.* Activation of the human c-kit product by ligand-induced dimerization mediates circular actin reorganization and chemotaxis. *EMBO J.* **10,** 4121–4128 (1991).
342. Roskoski, R. Signaling by Kit protein-tyrosine kinase--the stem cell factor receptor. *Biochem. Biophys. Res. Commun.* **337,** 1–13 (2005).
343. Kitamura, Y. & Hirotab, S. Kit as a human oncogenic tyrosine kinase. *Cell. Mol. Life Sci. CMLS* **61,** 2924–2931 (2004).
344. Roskoski, R. Structure and regulation of Kit protein-tyrosine kinase—The stem cell factor receptor. *Biochem. Biophys. Res. Commun.* **338,** 1307–1315 (2005).
345. Weiler, S. R. *et al.* JAK2 is associated with the c-kit proto-oncogene product and is phosphorylated in response to stem cell factor. *Blood* **87,** 3688–3693 (1996).
346. Valent, P., Sperr, W. R., Schwartz, L. B. & Horny, H.-P. Diagnosis and classification of mast cell proliferative disorders: delineation from immunologic diseases and non-mast cell hematopoietic neoplasms. *J. Allergy Clin. Immunol.* **114,** 3–11; quiz 12 (2004).
347. Arock, M. & Valent, P. Pathogenesis, classification and treatment of mastocytosis: state of the art in 2010 and future perspectives. *Expert Rev. Hematol.* **3,** 497–516 (2010).
348. Brockow, K. & Metcalfe, D. D. in *Chemical Immunology and Allergy* (ed. Ring, J.) **95,** 110–124 (KARGER, 2010).
349. Sánchez-Muñoz, L. *et al.* Evaluation of the WHO criteria for the classification of patients with mastocytosis. *Mod. Pathol. Off. J. U. S. Can. Acad. Pathol. Inc* **24,** 1157–1168 (2011).
350. Valent, P. Systemic mastocytosis. *Cancer Treat. Res.* **142,** 399–419 (2008).
351. Bibi, S. *et al.* Molecular defects in mastocytosis: KIT and beyond KIT. *Immunol. Allergy Clin. North Am.* **34,** 239–262 (2014).
352. Bodemer, C. *et al.* Pediatric mastocytosis is a clonal disease associated with D816V and other activating c-KIT mutations. *J. Invest. Dermatol.* **130,** 804–815 (2010).
353. Metcalfe, D. D. Mast cells and mastocytosis. *Blood* **112,** 946–956 (2008).
354. Tefferi, A., Li, C. Y., Butterfield, J. H. & Hoagland, H. C. Treatment of systemic mast-cell disease with cladribine. *N. Engl. J. Med.* **344,** 307–309 (2001).

355. Gleixner, K. V. *et al.* Synergistic growth-inhibitory effects of two tyrosine kinase inhibitors, dasatinib and PKC412, on neoplastic mast cells expressing the D816V-mutated oncogenic variant of KIT. *Haematologica* **92,** 1451–1459 (2007).
356. Akin, C. *et al.* Effects of tyrosine kinase inhibitor STI571 on human mast cells bearing wild-type or mutated c-kit. *Exp. Hematol.* **31,** 686–692 (2003).
357. Gotlib, J. *et al.* Midostaurin (PKC412) Demonstrates a High Rate of Durable Responses in Patients with Advanced Systemic Mastocytosis: Results from the Fully Accrued Global Phase 2 CPKC412D2201 Trial. at <https://ash.confex.com/ash/2014/webprogram/Paper67346.html>
358. Gleixner, K. V. *et al.* KIT-D816V-independent oncogenic signaling in neoplastic cells in systemic mastocytosis: role of Lyn and Btk activation and disruption by dasatinib and bosutinib. *Blood* **118,** 1885–1898 (2011).
359. Heltemes-Harris, L. M. *et al.* Ebf1 or Pax5 haploinsufficiency synergizes with STAT5 activation to initiate acute lymphoblastic leukemia. *J. Exp. Med.* **208,** 1135–1149 (2011).
360. Weber-Nordt, R. M. *et al.* Constitutive activation of STAT proteins in primary lymphoid and myeloid leukemia cells and in Epstein-Barr virus (EBV)-related lymphoma cell lines. *Blood* **88,** 809–816 (1996).
361. Nakayama, J. *et al.* BLNK suppresses pre-B-cell leukemogenesis through inhibition of JAK3. *Blood* **113,** 1483–1492 (2009).
362. Mullighan, C. G. *et al.* Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* **446,** 758–764 (2007).
363. Delogu, A. *et al.* Gene repression by Pax5 in B cells is essential for blood cell homeostasis and is reversed in plasma cells. *Immunity* **24,** 269–281 (2006).
364. Pongubala, J. M. R. *et al.* Transcription factor EBF restricts alternative lineage options and promotes B cell fate commitment independently of Pax5. *Nat. Immunol.* **9,** 203–215 (2008).
365. Kontro, M. *et al.* Novel activating STAT5B mutations as putative drivers of T-cell acute lymphoblastic leukemia. *Leukemia* **28,** 1738–1742 (2014).
366. Kuusanmäki, H. *et al.* Novel Activating STAT5B Mutations As Drivers Of T-ALL. *Blood* **122,** 3863–3863 (2013).
367. Sallmyr, A. *et al.* Internal tandem duplication of FLT3 (FLT3/ITD) induces increased ROS production, DNA damage, and misrepair: implications for poor prognosis in AML. *Blood* **111,** 3173–3182 (2008).
368. Woolley, J. F. *et al.* H2O2 production downstream of FLT3 is mediated by p22phox in the endoplasmic reticulum and is required for STAT5 signalling. *PloS One* **7,** e34050 (2012).
369. Cotarla, I. *et al.* Stat5a is tyrosine phosphorylated and nuclear localized in a high proportion of human breast cancers. *Int. J. Cancer J. Int. Cancer* **108,** 665–671 (2004).
370. Walker, S. R. *et al.* Reciprocal effects of STAT5 and STAT3 in breast cancer. *Mol. Cancer Res. MCR* **7,** 966–976 (2009).
371. Iavnilovitch, E., Groner, B. & Barash, I. Overexpression and forced activation of stat5 in mammary gland of transgenic mice promotes cellular proliferation, enhances differentiation, and delays postlactational apoptosis. *Mol. Cancer Res. MCR* **1,** 32–47 (2002).
372. Nevalainen, M. T. *et al.* Signal transducer and activator of transcription-5 activation and breast cancer prognosis. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **22,** 2053–2060 (2004).
373. Nouhi, Z. *et al.* Defining the role of prolactin as an invasion suppressor hormone in breast cancer cells. *Cancer Res.* **66,** 1824–1832 (2006).
374. Sultan, A. S. *et al.* Stat5 promotes homotypic adhesion and inhibits invasive characteristics of human breast cancer cells. *Oncogene* **24,** 746–760 (2005).
375. Van de Vijver, M. J. *et al.* A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347,** 1999–2009 (2002).
376. Ni Chonghaile, T. *et al.* Pretreatment mitochondrial priming correlates with clinical response to cytotoxic chemotherapy. *Science* **334,** 1129–1133 (2011).

377. Ryan, J. A., Brunelle, J. K. & Letai, A. Heightened mitochondrial priming is the basis for apoptotic hypersensitivity of CD4+ CD8+ thymocytes. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 12895–12900 (2010).
378. Barclay, J. L., Anderson, S. T., Waters, M. J. & Curlewis, J. D. SOCS3 as a tumor suppressor in breast cancer cells, and its regulation by PRL. *Int. J. Cancer J. Int. Cancer* **124,** 1756–1766 (2009).
379. Matsumoto, A. *et al.* CIS, a cytokine inducible SH2 protein, is a target of the JAK-STAT5 pathway and modulates STAT5 activation. *Blood* **89,** 3148–3154 (1997).
380. Sutherland, K. D., Lindeman, G. J. & Visvader, J. E. Knocking off SOCS genes in the mammary gland. *Cell Cycle Georget. Tex* **6,** 799–803 (2007).
381. Ahonen, T. J. *et al.* Inhibition of transcription factor Stat5 induces cell death of human prostate cancer cells. *J. Biol. Chem.* **278,** 27287–27292 (2003).
382. Li, H. *et al.* Activation of signal transducer and activator of transcription 5 in human prostate cancer is associated with high histological grade. *Cancer Res.* **64,** 4774–4782 (2004).
383. Haddad, B. R. *et al.* STAT5A/B Gene Locus Undergoes Amplification during Human Prostate Cancer Progression. *Am. J. Pathol.* **182,** 2264–2275 (2013).
384. Ahonen, T. J., Härkönen, P. L., Rui, H. & Nevalainen, M. T. PRL signal transduction in the epithelial compartment of rat prostate maintained as long-term organ cultures in vitro. *Endocrinology* **143,** 228–238 (2002).
385. Dagvadorj, A. *et al.* Autocrine prolactin promotes prostate cancer cell growth via Janus kinase-2-signal transducer and activator of transcription-5a/b signaling pathway. *Endocrinology* **148,** 3089–3101 (2007).
386. Nevalainen, M. T. *et al.* Prolactin and prolactin receptors are expressed and functioning in human prostate. *J. Clin. Invest.* **99,** 618–627 (1997).
387. Dagvadorj, A., Kirken, R. A., Leiby, B., Karras, J. & Nevalainen, M. T. Transcription factor signal transducer and activator of transcription 5 promotes growth of human prostate cancer cells in vivo. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **14,** 1317–1324 (2008).
388. Sackmann-Sala, L. & Goffin, V. Prolactin-induced prostate tumorigenesis. *Adv. Exp. Med. Biol.* **846,** 221–242 (2015).
389. Sackmann-Sala, L. *et al.* Prolactin-induced prostate tumorigenesis links sustained Stat5 signaling with the amplification of basal/stem cells and emergence of putative luminal progenitors. *Am. J. Pathol.* **184,** 3105–3119 (2014).
390. Gu, L. *et al.* Stat5 promotes metastatic behavior of human prostate cancer cells in vitro and in vivo. *Endocr. Relat. Cancer* **17,** 481–493 (2010).
391. Li, H. *et al.* Activation of signal transducer and activator of transcription-5 in prostate cancer predicts early recurrence. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **11,** 5863–5868 (2005).
392. Tan, S.-H. *et al.* Transcription factor Stat5 synergizes with androgen receptor in prostate cancer cells. *Cancer Res.* **68,** 236–248 (2008).
393. Thomas, C. *et al.* Transcription factor Stat5 knockdown enhances androgen receptor degradation and delays castration-resistant prostate cancer progression in vivo. *Mol. Cancer Ther.* **10,** 347–359 (2011).
394. Mirtti, T. *et al.* Nuclear Stat5a/b predicts early recurrence and prostate cancer-specific death in patients treated by radical prostatectomy. *Hum. Pathol.* **44,** 310–319 (2013).
395. Loughran, T. P. *et al.* Leukemia of large granular lymphocytes: association with clonal chromosomal abnormalities and autoimmune neutropenia, thrombocytopenia, and hemolytic anemia. *Ann. Intern. Med.* **102,** 169–175 (1985).
396. Sokol, L. & Loughran, T. P. Large granular lymphocyte leukemia. *The Oncologist* **11,** 263–273 (2006).

397. Jerez, A. *et al.* STAT3 mutations unify the pathogenesis of chronic lymphoproliferative disorders of NK cells and T-cell large granular lymphocyte leukemia. *Blood* **120,** 3048–3057 (2012).
398. Koskela, H. L. M. *et al.* Somatic *STAT3* Mutations in Large Granular Lymphocytic Leukemia. *N. Engl. J. Med.* **366,** 1905–1913 (2012).
399. Andersson, E. I. *et al.* Novel somatic mutations in large granular lymphocytic leukemia affecting the STAT-pathway and T-cell activation. *Blood Cancer J* **3,** e168 (2013).
400. Rajala, H. L. *et al.* Discovery of somatic STAT5b mutations in large granular lymphocytic leukemia. *Blood* **121,** 4541–4550 (2013).
401. Hopfinger, G. *et al.* Sequential chemoimmunotherapy of fludarabine, mitoxantrone, and cyclophosphamide induction followed by alemtuzumab consolidation is effective in T-cell prolymphocytic leukemia: FMC Plus Alemtuzumab in T-PLL. *Cancer* **119,** 2258–2267 (2013).
402. Matutes, E. *et al.* Clinical and laboratory features of 78 cases of T-prolymphocytic leukemia. *Blood* **78,** 3269–3274 (1991).
403. Dearden, C. E. T-cell prolymphocytic leukemia. *Med. Oncol. Northwood Lond. Engl.* **23,** 17–22 (2006).
404. Stilgenbauer, S. *et al.* Biallelic mutations in the ATM gene in T-prolymphocytic leukemia. *Nat. Med.* **3,** 1155–1159 (1997).
405. Kiel, M. J. *et al.* Integrated genomic sequencing reveals mutational landscape of T-cell prolymphocytic leukemia. *Blood* **124,** 1460–1472 (2014).
406. Mow, B. M. F. *et al.* Effects of the Bcr/abl kinase inhibitors STI571 and adaphostin (NSC 680410) on chronic myelogenous leukemia cells in vitro. *Blood* **99,** 664–671 (2002).
407. Huang, M. *et al.* Inhibition of Bcr-Abl kinase activity by PD180970 blocks constitutive activation of Stat5 and growth of CML cells. *Oncogene* **21,** 8804–8816 (2002).
408. Kelly, L. M. *et al.* FLT3 internal tandem duplication mutations associated with human acute myeloid leukemias induce myeloproliferative disease in a murine bone marrow transplant model. *Blood* **99,** 310–318 (2002).
409. Rocnik, J. L. *et al.* Roles of tyrosine 589 and 591 in STAT5 activation and transformation mediated by FLT3-ITD. *Blood* **108,** 1339–1345 (2006).
410. Auclair, D. *et al.* Antitumor activity of sorafenib in FLT3-driven leukemic cells. *Leukemia* **21,** 439–445 (2007).
411. Chao, Q. *et al.* Identification of N-(5-tert-butyl-isoxazol-3-yl)-N′-{4-[7-(2-morpholin-4-yl-ethoxy)imidazo[2,1-b][1,3]benzothiazol-2-yl]phenyl}urea dihydrochloride (AC220), a uniquely potent, selective, and efficacious FMS-like tyrosine kinase-3 (FLT3) inhibitor. *J. Med. Chem.* **52,** 7808–7816 (2009).
412. Levis, M. *et al.* A FLT3-targeted tyrosine kinase inhibitor is cytotoxic to leukemia cells in vitro and in vivo. *Blood* **99,** 3885–3891 (2002).
413. Sato, T. *et al.* FLT3 ligand impedes the efficacy of FLT3 inhibitors in vitro and in vivo. *Blood* **117,** 3286–3293 (2011).
414. Pardanani, A. *et al.* CYT387, a selective JAK1/JAK2 inhibitor: in vitro assessment of kinase selectivity and preclinical studies using cell lines and primary cells from polycythemia vera patients. *Leukemia* **23,** 1441–1445 (2009).
415. Quintás-Cardama, A. *et al.* Preclinical characterization of the selective JAK1/2 inhibitor INCB018424: therapeutic implications for the treatment of myeloproliferative neoplasms. *Blood* **115,** 3109–3117 (2010).
416. Santos, F. P. S. *et al.* Phase 2 study of CEP-701, an orally available JAK2 inhibitor, in patients with primary or post-polycythemia vera/essential thrombocythemia myelofibrosis. *Blood* **115,** 1131–1136 (2010).
417. Wernig, G. *et al.* Efficacy of TG101348, a selective JAK2 inhibitor, in treatment of a murine model of JAK2V617F-induced polycythemia vera. *Cancer Cell* **13,** 311–320 (2008).

418. Schust, J., Sperl, B., Hollis, A., Mayer, T. U. & Berg, T. Stattic: a small-molecule inhibitor of STAT3 activation and dimerization. *Chem. Biol.* **13,** 1235–1242 (2006).
419. Müller, J., Sperl, B., Reindl, W., Kiessling, A. & Berg, T. Discovery of Chromone-Based Inhibitors of the Transcription Factor STAT5. *ChemBioChem* **9,** 723–727 (2008).
420. Siddiquee, K. *et al.* Selective chemical probe inhibitor of Stat3, identified through structure-based virtual screening, induces antitumor activity. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 7391–7396 (2007).
421. Page, B. D. *et al.* Small molecule STAT5-SH2 domain inhibitors exhibit potent antileukemia activity. *J.Med.Chem.* **55,** 1047–1055 (2012).
422. Nelson, E. A. *et al.* The STAT5 inhibitor pimozide decreases survival of chronic myelogenous leukemia cells resistant to kinase inhibitors. *Blood* **117,** 3421–3429 (2011).
423. Nam, S. *et al.* Indirubin derivatives induce apoptosis of chronic myelogenous leukemia cells involving inhibition of Stat5 signaling. *Mol.Oncol.* **6,** 276–283 (2012).
424. Hayakawa, F. *et al.* A novel STAT inhibitor, OPB-31121, has a significant antitumor effect on leukemia with STAT-addictive oncokinases. *Blood Cancer J.* **3,** e166 (2013).
425. Cumaraswamy, A. A. *et al.* Nanomolar-Potency Small Molecule Inhibitor of STAT5 Protein. *ACS Med. Chem. Lett.* **5,** 1202–1206 (2014).
426. Wang, X. *et al.* Targeted blockage of signal transducer and activator of transcription 5 signaling pathway with decoy oligodeoxynucleotides suppresses leukemic K562 cell growth. *DNA Cell Biol.* **30,** 71–78 (2011).
427. Behbod, F. *et al.* Specific inhibition of Stat5a/b promotes apoptosis of IL-2-responsive primary and tumor-derived lymphoid cells. *J. Immunol. Baltim. Md 1950* **171,** 3919–3927 (2003).
428. Weber, A. *et al.* The inhibition of stat5 by a Peptide aptamer ligand specific for the DNA binding domain prevents target gene transactivation and the growth of breast and prostate tumor cells. *Pharmaceuticals.(Basel)* **6,** 960–987 (2013).
429. Liu, S. *et al.* Targeting STAT5 in hematologic malignancies through inhibition of the bromodomain and extra-terminal (BET) bromodomain protein BRD2. *Mol. Cancer Ther.* **13,** 1194–1205 (2014).
430. Timofeeva, O. A. & Tarasova, N. I. Alternative ways of modulating JAK-STAT pathway: Looking beyond phosphorylation. *JAK-STAT* **1,** 274–284 (2012).
431. Berger, A., Sexl, V., Valent, P. & Moriggl, R. Inhibition of STAT5: a therapeutic option in BCR-ABL1-driven leukemia. *Oncotarget* **5,** 9564–9576 (2014).
432. Williams, D. H., Searle, M. S., Mackay, J. P., Gerhard, U. & Maplestone, R. A. Toward an estimation of binding constants in aqueous solution: studies of associations of vancomycin group antibiotics. *Proc. Natl. Acad. Sci.* **90,** 1172–1178 (1993).
433. Davis, A. M. & Teague, S. J. Hydrogen Bonding, Hydrophobic Interactions, and Failure of the Rigid Receptor Hypothesis. *Angew. Chem. Int. Ed.* **38,** 736–749 (1999).
434. Creighton, T. E. Disulphide bonds and protein stability. *BioEssays* **8,** 57–63 (1988).
435. Thornton, J. M. Disulphide bridges in globular proteins. *J. Mol. Biol.* **151,** 261–287 (1981).
436. Katz, B. A. & Kossiakoff, A. The crystallographically determined structures of atypical strained disulfides engineered into subtilisin. *J. Biol. Chem.* **261,** 15480–15485 (1986).
437. Pace, C. N. *et al.* Contribution of Hydrophobic Interactions to Protein Stability. *J. Mol. Biol.* **408,** 514–528 (2011).
438. Kendrew, J. C. *et al.* A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature* **181,** 662–666 (1958).
439. Schmidt, A., Teeter, M., Weckert, E. & Lamzin, V. S. Crystal structure of small protein crambin at 0.48 Å resolution. *Acta Crystallograph. Sect. F Struct. Biol. Cryst. Commun.* **67,** 424–428 (2011).

440. Polikanov, Y. S., Steitz, T. A. & Innis, C. A. A proton wire to couple aminoacyl-tRNA accommodation and peptide-bond formation on the ribosome. *Nat. Struct. Mol. Biol.* **21,** 787–793 (2014).
441. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).
442. DeLano, W. *The PyMOL Molecular Graphics System, v1.7*. (Schrödinger, LLC).
443. *Schrödinger Release 2015-2: Maestro*. (Schrödinger, LLC, New York, NY, 2015).
444. Viani, L. *MView: A tool for visualization and analysis of molecular properties.* at <www.mview-tools.com>
445. Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).
446. Humphrey, W., Dalke, A. & Schulten, K. VMD – Visual Molecular Dynamics. *J. Mol. Graph.* **14,** 33–38 (1996).
447. Beuming, T. *et al.* The binding sites for cocaine and dopamine in the dopamine transporter overlap. *Nat. Neurosci.* **11,** 780–789 (2008).
448. Lewis, D. The CAVE artists. *Nat. Med.* **20,** 228–230 (2014).
449. McIntosh *et al.* Computing power revolution and new algorithms: GP-GPUs, clouds and more: general discussion. *Faraday Discuss* **169,** 379–401 (2014).
450. Sela, M., White, F. H. & Anfinsen, C. B. Reductive Cleavage of Disulfide Bridges in Ribonuclease. *Science* **125,** 691–692 (1957).
451. Anfinsen, C. B. Principles that Govern the Folding of Protein Chains. *Science* **181,** 223–230 (1973).
452. Pace, C. N., Shirley, B. A., McNutt, M. & Gajiwala, K. Forces contributing to the conformational stability of proteins. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.* **10,** 75–83 (1996).
453. Van den Berg, B., Wain, R., Dobson, C. M. & Ellis, R. J. Macromolecular crowding perturbs protein refolding kinetics: implications for folding inside the cell. *EMBO J.* **19,** 3870–3875 (2000).
454. Huber, R. & Bennett, W. S. Functional significance of flexibility in proteins. *Biopolymers* **22,** 261–279 (1983).
455. Chothia, C. & Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5,** 823–826 (1986).
456. Zhang, Y. & Skolnick, J. The protein structure prediction problem could be solved using the current PDB library. *Proc. Natl. Acad. Sci.* **102,** 1029–1034 (2005).
457. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
458. Lipman, D. & Pearson, W. Rapid and sensitive protein similarity searches. *Science* **227,** 1435–1441 (1985).
459. Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.* **85,** 2444–2448 (1988).
460. Baker, D. & Sali, A. Protein Structure Prediction and Structural Genomics. *Science* **294,** 93–96 (2001).
461. Fiser, A., Do, R. K. & Sali, A. Modeling of loops in protein structures. *Protein Sci. Publ. Protein Soc.* **9,** 1753–1773 (2000).
462. Lathrop, R. H. The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein Eng.* **7,** 1059–1068 (1994).
463. Westhead, D. R. *et al.* Protein fold recognition by threading: comparison of algorithms and analysis of results. *Protein Eng.* **8,** 1197–1204 (1995).
464. Chothia, C. Proteins. One thousand families for the molecular biologist. *Nature* **357,** 543–544 (1992).
465. Leonov, H., Mitchell, J. S. B. & Arkin, I. T. Monte Carlo estimation of the number of possible protein folds: effects of sampling bias and folds distributions. *Proteins* **51,** 352–359 (2003).

466. Englander, S. W. & Mayne, L. The nature of protein folding pathways. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 15873–15880 (2014).
467. Floudas, C. A. Computational methods in protein structure prediction. *Biotechnol. Bioeng.* **97,** 207–213 (2007).
468. Bonneau, R. & Baker, D. Ab initio protein structure prediction: progress and prospects. *Annu. Rev. Biophys. Biomol. Struct.* **30,** 173–189 (2001).
469. Breda, A., Santos, D. S., Basso, L. A. & de Souza, O. N. Ab initio 3-D structure prediction of an artificially designed three-alpha-helix bundle via all-atom molecular dynamics simulations. *Genet. Mol. Res. GMR* **6,** 901–910 (2007).
470. Díaz, N. & Suárez, D. Extensive Simulations of the Full-Length Matrix Metalloproteinase-2 Enzyme in a Prereactive Complex with a Collagen Triple-Helical Peptide. *Biochemistry (Mosc.)* 150128143332005 (2015). doi:10.1021/bi501014w
471. Koldsø, H., Shorthouse, D., Hélie, J. & Sansom, M. S. P. Lipid Clustering Correlates with Membrane Curvature as Revealed by Molecular Simulations of Complex Lipid Bilayers. *PLoS Comput. Biol.* **10,** e1003911 (2014).
472. Zhao, G. *et al.* Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature* **497,** 643–646 (2013).
473. Shaw, D. E. *et al.* Anton, a special-purpose machine for molecular dynamics simulation. *Commun. ACM* **51,** 91 (2008).
474. Piana, S., Lindorff-Larsen, K. & Shaw, D. E. Atomic-level description of ubiquitin folding. *Proc. Natl. Acad. Sci.* **110,** 5915–5920 (2013).
475. Nyquist, H. Certain Topics in Telegraph Transmission Theory. *Trans. Am. Inst. Electr. Eng.* **47,** 617–644 (1928).
476. Hockney, R. ., Goel, S. . & Eastwood, J. . Quiet high-resolution computer models of a plasma. *J. Comput. Phys.* **14,** 148–158 (1974).
477. Swope, W. C. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *J. Chem. Phys.* **76,** 637 (1982).
478. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N [center-dot] log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98,** 10089–10092 (1993).
479. Ewald, P. P. Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Ann. Phys.* **369,** 253–287 (1921).
480. Laio, A. & Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci.* **99,** 12562–12566 (2002).
481. Schlitter, J., Engels, M. & Krüger, P. Targeted molecular dynamics: a new approach for searching pathways of conformational transitions. *J. Mol. Graph.* **12,** 84–89 (1994).
482. Grubmüller, H., Heymann, B. & Tavan, P. Ligand binding: molecular mechanics calculation of the streptavidin-biotin rupture force. *Science* **271,** 997–999 (1996).
483. Isralewitz, B., Izrailev, S. & Schulten, K. Binding pathway of retinal to bacterio-opsin: a prediction by molecular dynamics simulations. *Biophys. J.* **73,** 2972–2979 (1997).
484. Hinsen, K., Petrescu, A.-J., Dellerue, S., Bellissent-Funel, M.-C. & Kneller, G. R. Harmonicity in slow protein dynamics. *Chem. Phys.* **261,** 25–37 (2000).
485. Bahar, I., Lezon, T. R., Bakan, A. & Shrivastava, I. H. Normal Mode Analysis of Biomolecular Structures: Functional Mechanisms of Membrane Proteins. *Chem. Rev.* **110,** 1463–1497 (2010).
486. Al-Bluwi, I., Vaisset, M., Siméon, T. & Cortés, J. Modeling protein conformational transitions by a combination of coarse-grained normal mode analysis and robotics-inspired methods. *BMC Struct. Biol.* **13,** S2 (2013).
487. Mouawad, L. & Perahia, D. Diagonalization in a mixed basis: A method to compute low-frequency normal modes for large macromolecules. *Biopolymers* **33,** 599–611 (1993).

488. Marques, O. & Sanejouand, Y.-H. Hinge-bending motion in citrate synthase arising from normal mode calculations. *Proteins Struct. Funct. Genet.* **23,** 557–560 (1995).
489. Perahia, D. & Mouawad, L. Computation of low-frequency normal modes in macromolecules: improvements to the method of diagonalization in a mixed basis and application to hemoglobin. *Comput. Chem.* **19,** 241–246 (1995).
490. Durand, P., Trinquier, G. & Sanejouand, Y.-H. A new approach for determining low-frequency normal modes in macromolecules. *Biopolymers* **34,** 759–771 (1994).
491. Tama, F., Gadea, F. X., Marques, O. & Sanejouand, Y. H. Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins* **41,** 1–7 (2000).
492. Tirion, null. Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Phys. Rev. Lett.* **77,** 1905–1908 (1996).
493. Bahar, I., Atilgan, A. R. & Erman, B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.* **2,** 173–181 (1997).
494. Atilgan, A. R. *et al.* Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model. *Biophys. J.* **80,** 505–515 (2001).
495. Monod, J. *Chance and necessity; an essay on the natural philosophy of modern biology*. (Knopf, 1971).
496. Changeux, J.-P. Allostery and the Monod-Wyman-Changeux Model After 50 Years. *Annu. Rev. Biophys.* **41,** 103–133 (2012).
497. Changeux, J.-P. 50 years of allosteric interactions: the twists and turns of the models. *Nat. Rev. Mol. Cell Biol.* **14,** 819–829 (2013).
498. Vendruscolo, M. Protein regulation: The statistical theory of allostery. *Nat. Chem. Biol.* **7,** 411–412 (2011).
499. Bohr, C., Hasselbalch, K. & Krogh, A. Ueber einen in biologischer Beziehung wichtigen Einfluss, den die Kohlensäurespannung des Blutes auf dessen Sauerstoffbindung übt1. *Skand. Arch. Für Physiol.* **16,** 402–412 (1904).
500. Hill, A. V. The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curve. *J Physiol* **40,** (1910).
501. Changeux, J. P., Podleski, T. R. & Wofsy, L. Affinity labeling of the acetylcholine-receptor. *Proc. Natl. Acad. Sci. U. S. A.* **58,** 2063–2070 (1967).
502. Monod, J., Wyman, J. & Changeux, J.-P. On the nature of allosteric transitions: A plausible model. *J. Mol. Biol.* **12,** 88–118 (1965).
503. Koshland, D. E., Némethy, G. & Filmer, D. Comparison of Experimental Binding Data and Theoretical Models in Proteins Containing Subunits *. *Biochemistry (Mosc.)* **5,** 365–385 (1966).
504. Palczewski, K. *et al.* Crystal structure of rhodopsin: A G protein-coupled receptor. *Science* **289,** 739–745 (2000).
505. Sobolevsky, A. I., Rosconi, M. P. & Gouaux, E. X-ray structure, symmetry and mechanism of an AMPA-subtype glutamate receptor. *Nature* **462,** 745–756 (2009).
506. Colombo, M., Rau, D. & Parsegian, V. Protein solvation in allosteric regulation: a water effect on hemoglobin. *Science* **256,** 655–659 (1992).
507. Cooper, A. & Dryden, D. T. F. Allostery without conformational change: A plausible model. *Eur. Biophys. J.* **11,** 103–109 (1984).
508. Jaffe, E. K. Morpheeins--a new structural paradigm for allosteric regulation. *Trends Biochem. Sci.* **30,** 490–497 (2005).
509. Motlagh, H. N., Wrabl, J. O., Li, J. & Hilser, V. J. The ensemble nature of allostery. *Nature* **508,** 331–339 (2014).
510. Luque, I., Leavitt, S. A. & Freire, E. T HE L INKAGE B ETWEEN P ROTEIN F OLDING AND F UNCTIONAL C OOPERATIVITY : Two Sides of the Same Coin? *Annu. Rev. Biophys. Biomol. Struct.* **31,** 235–256 (2002).

511. Hilser, V. J. & Thompson, E. B. Intrinsic disorder as a mechanism to optimize allosteric coupling in proteins. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 8311–8315 (2007).
512. Hilser, V. J., Wrabl, J. O. & Motlagh, H. N. Structural and Energetic Basis of Allostery. *Annu. Rev. Biophys.* **41,** 585–609 (2012).
513. Clarkson, M. W. & Lee, A. L. Long-range dynamic effects of point mutations propagate through side chains in the serine protease inhibitor eglin c. *Biochemistry (Mosc.)* **43,** 12448–12458 (2004).
514. Schrank, T. P., Bolen, D. W. & Hilser, V. J. Rational modulation of conformational fluctuations in adenylate kinase reveals a local unfolding mechanism for allostery and functional adaptation in proteins. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 16984–16989 (2009).
515. Dixit, A. & Verkhivker, G. M. Hierarchical modeling of activation mechanisms in the ABL and EGFR kinase domains: thermodynamic and mechanistic catalysts of kinase activation by cancer mutations. *PLoS Comput. Biol.* **5,** e1000487 (2009).
516. Daily, M. D. & Gray, J. J. Allosteric Communication Occurs via Networks of Tertiary and Quaternary Motions in Proteins. *PLoS Comput. Biol.* **5,** e1000293 (2009).
517. Kazi, J. U., Kabir, N. N., Flores-Morales, A. & Rönnstrand, L. SOCS proteins in regulation of receptor tyrosine kinase signaling. *Cell. Mol. Life Sci. CMLS* **71,** 3297–3310 (2014).
518. Nussinov, R., Tsai, C.-J. & Ma, B. The Underappreciated Role of Allostery in the Cellular Network. *Annu. Rev. Biophys.* **42,** 169–189 (2013).
519. Kalatskaya, I. *et al.* AMD3100 Is a CXCR7 Ligand with Allosteric Agonist Properties. *Mol. Pharmacol.* **75,** 1240–1247 (2009).
520. Redka, D. S., Pisterzi, L. F. & Wells, J. W. Binding of Orthosteric Ligands to the Allosteric Site of the M2 Muscarinic Cholinergic Receptor. *Mol. Pharmacol.* **74,** 834–843 (2008).
521. Maillet, E. L. *et al.* A novel, conformation-specific allosteric inhibitor of the tachykinin NK2 receptor (NK2R) with functionally selective properties. *FASEB J.* **21,** 2124–2134 (2007).
522. Gatson, J. W., Simpkins, J. W. & Uteshev, V. V. High therapeutic potential of positive allosteric modulation of α7 nAChRs in a rat model of traumatic brain injury: Proof-of-concept. *Brain Res. Bull.* (2015). doi:10.1016/j.brainresbull.2015.01.008
523. Laine, E., Chauvot de Beauchêne, I., Perahia, D., Auclair, C. & Tchertanov, L. Mutation D816V Alters the Internal Structure and Dynamics of c-KIT Receptor Cytoplasmic Region: Implications for Dimerization and Activation Mechanisms. *PLoS Comput. Biol.* **7,** e1002068 (2011).
524. Soldaini, E. *et al.* DNA binding site selection of dimeric and tetrameric Stat5 proteins reveals a large repertoire of divergent tetrameric Stat5a binding sites. *MolCell Biol* **20,** 389–401 (2000).
525. Razeto, A. *et al.* Structure of the NCoA-1/SRC-1 PAS-B domain bound to the LXXLL motif of the STAT6 transactivation domain. *J. Mol. Biol.* **336,** 319–329 (2004).
526. Wojciak, J. M., Martinez-Yamout, M. A., Dyson, H. J. & Wright, P. E. Structural basis for recruitment of CBP/p300 coactivators by STAT1 and STAT2 transactivation domains. *EMBO J.* **28,** 948–958 (2009).
527. Moriggl, R. *et al.* Deletion of the carboxyl-terminal transactivation domain of MGF-Stat5 results in sustained DNA binding and a dominant negative phenotype. *Mol. Cell. Biol.* **16,** 5691–5700 (1996).
528. Schindler, C. & Darnell, J. E. Transcriptional responses to polypeptide ligands: the JAK-STAT pathway. *Annu. Rev. Biochem.* **64,** 621–651 (1995).
529. Horvath, C. M., Wen, Z. & Darnell, J. E., Jr. A STAT protein domain that determines DNA sequence recognition suggests a novel DNA-binding domain. *Genes Dev* **9,** 984–994 (1995).
530. Ehret, G. B. *et al.* DNA binding specificity of different STAT proteins. Comparison of in vitro specificity with natural target sites. *J.Biol.Chem.* **276,** 6675–6688 (2001).

531. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of *Coot*. *Acta Crystallogr. D Biol. Crystallogr.* **66,** 486–501 (2010).
532. Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J.Mol.Biol.* **234,** 779–815 (1993).
533. Shen, M. Y. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci* **15,** 2507–2524 (2006).
534. Laskowski, R. A., Rullmannn, J. A., MacArthur, M. W., Kaptein, R. & Thornton, J. M. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR* **8,** 477–486 (1996).
535. Gordon, J. C. *et al.* H++: a server for estimating pKas and adding missing hydrogens to macromolecules. *Nucleic Acids Res* **33,** W368–W371 (2005).
536. Myers, J., Grothaus, G., Narayanan, S. & Onufriev, A. A simple clustering algorithm can be accurate enough for use in calculations of pKs in macromolecules. *Proteins* **63,** 928–938 (2006).
537. Anandakrishnan, R., Aguilar, B. & Onufriev, A. V. H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Res* **40,** W537–W541 (2012).
538. Husby, J. *et al.* Molecular dynamics studies of the STAT3 homodimer:DNA complex: relationships between STAT3 mutations and protein-DNA recognition. *J.Chem.Inf.Model.* **52,** 1179–1192 (2012).
539. Pronk, S. *et al.* GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics.* (2013). doi:10.1093/bioinformatics/btt055
540. Hornak, V. *et al.* Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **65,** 712–725 (2006).
541. Best, R. B. & Hummer, G. Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. *J.Phys.Chem.B* **113,** 9004–9015 (2009).
542. Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78,** 1950–1958 (2010).
543. Homeyer, N., Horn, A. H., Lanig, H. & Sticht, H. AMBER force-field parameters for phosphorylated amino acids in different protonation states: phosphoserine, phosphothreonine, phosphotyrosine, and phosphohistidine. *J.Mol.Model.* **12,** 281–289 (2006).
544. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79,** 926 (1983).
545. Berendsen, H. J. C., Grigera, J. R. & Straatsma, T. P. The missing term in effective pair potentials. *J. Phys. Chem.* **91,** 6269–6271 (1987).
546. Bernal, J. D. & Fowler, R. H. A Theory of Water and Ionic Solution, with Particular Reference to Hydrogen and Hydroxyl Ions. *J. Chem. Phys.* **1,** 515 (1933).
547. Mahoney, M. W. & Jorgensen, W. L. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J. Chem. Phys.* **112,** 8910 (2000).
548. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81,** 3684 (1984).
549. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126,** 014101 (2007).
550. Parrinello, M. & Rahman, K. M. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **52,** 7182 (1981).
551. Ohno, I., Tomizawa, M., Durkin, K. A., Casida, J. E. & Kagabu, S. Neonicotinoid Substituents Forming a Water Bridge at the Nicotinic Acetylcholine Receptor. *J. Agric. Food Chem.* **57,** 2436–2440 (2009).

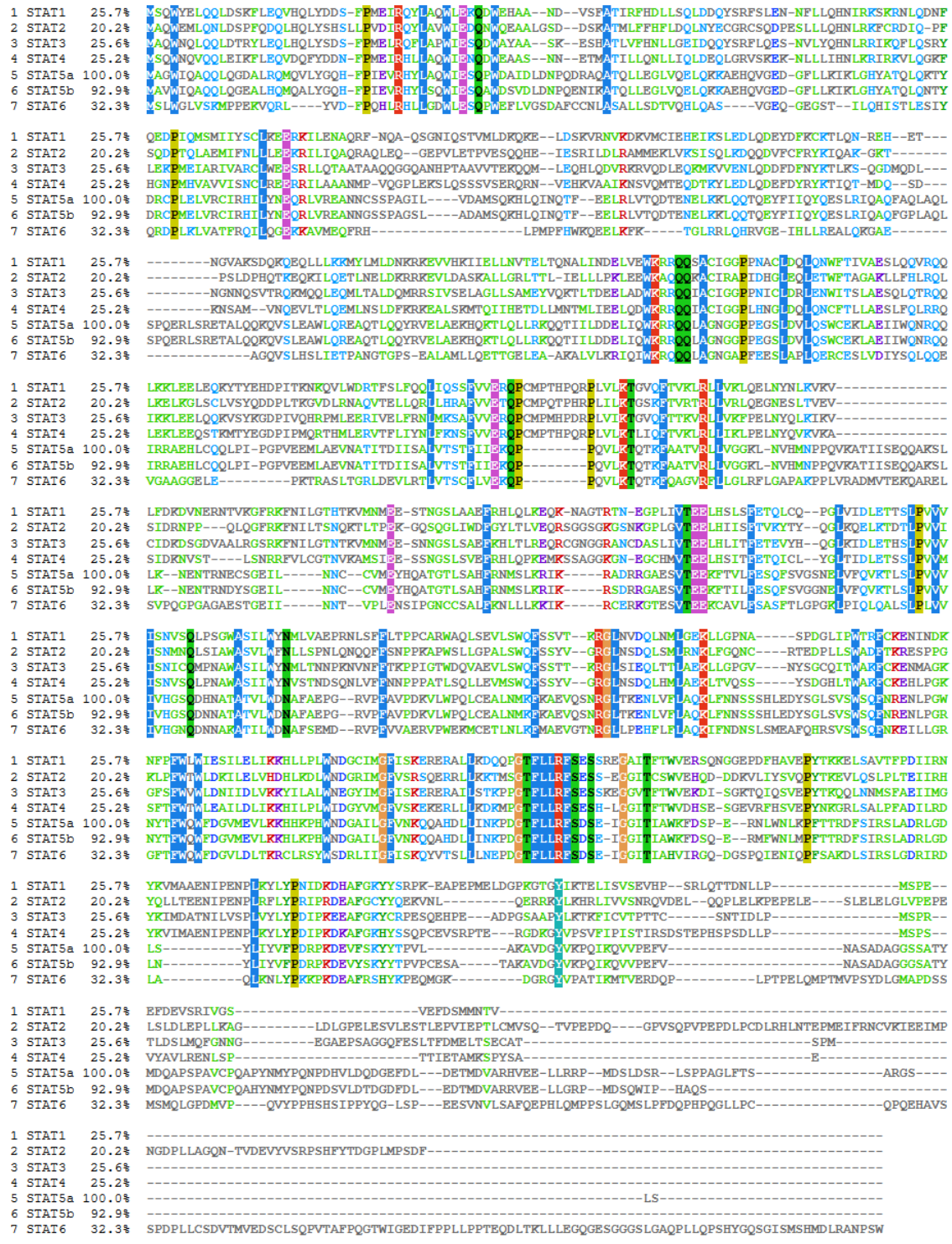
552. Sun, X., Ågren, H. & Tu, Y. Functional Water Molecules in Rhodopsin Activation. *J. Phys. Chem. B* **118,** 10863–10873 (2014).
553. Xu, S. *et al.* Oxicams Bind in a Novel Mode to the Cyclooxygenase Active Site via a Two-water-mediated H-bonding Network. *J. Biol. Chem.* **289,** 6799–6808 (2014).
554. Dahl, A. C. E., Chavent, M. & Sansom, M. S. P. Bendix: intuitive helix geometry analysis and abstraction. *Bioinformatics* **28,** 2193–2194 (2012).
555. Bakan, A., Meireles, L. M. & Bahar, I. ProDy: protein dynamics inferred from theory and experiments. *Bioinformatics.* **27,** 1575–1577 (2011).
556. Amadei, A., Linssen, A. B. & Berendsen, H. J. Essential dynamics of proteins. *Proteins* **17,** 412–425 (1993).
557. Eyal, E., Chennubhotla, C., Yang, L. W. & Bahar, I. Anisotropic fluctuations of amino acids in protein structures: insights from X-ray crystallography and elastic network models. *Bioinformatics.* **23,** i175–i184 (2007).
558. Mccammon, J. A., Gelin, B. R., Karplus, M. & Wolynes, P. G. The hinge-bending mode in lysozyme. *Nature* **262,** 325–326 (1976).
559. Lezon, T. R. & Bahar, I. Using Entropy Maximization to Understand the Determinants of Structural Dynamics beyond Native Contact Topology. *PLoS Comput. Biol.* **6,** e1000816 (2010).
560. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22,** 2577–2637 (1983).
561. Joosten, R. P. *et al.* A series of PDB related databases for everyday needs. *Nucleic Acids Res.* **39,** D411–D419 (2011).
562. Atilgan, A. R., Turgut, D. & Atilgan, C. Screened Nonbonded Interactions in Native Proteins Manipulate Optimal Paths for Robust Residue Communication. *Biophys. J.* **92,** 3052–3062 (2007).
563. Chennubhotla, C. & Bahar, I. Signal propagation in proteins and relation to equilibrium fluctuations. *PLoS.Comput.Biol.* **3,** 1716–1726 (2007).
564. Park, K. & Kim, D. Modeling allosteric signal propagation using protein structure networks. *BMC Bioinformatics* **12,** S23 (2011).
565. Okazaki, K. -i., Koga, N., Takada, S., Onuchic, J. N. & Wolynes, P. G. Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: Structure-based molecular dynamics simulations. *Proc. Natl. Acad. Sci.* **103,** 11844–11849 (2006).
566. Sethi, A., Eargle, J., Black, A. A. & Luthey-Schulten, Z. Dynamical networks in tRNA:protein complexes. *Proc. Natl. Acad. Sci.* **106,** 6620–6625 (2009).
567. Law, S. M., Gagnon, J. K., Mapp, A. K. & Brooks, C. L. Prepaying the entropic cost for allosteric regulation in KIX. *Proc. Natl. Acad. Sci.* **111,** 12067–12072 (2014).
568. Laine, E., Auclair, C. & Tchertanov, L. Allosteric communication across the native and mutated KIT receptor tyrosine kinase. *PLoS.Comput.Biol.* **8,** e1002661 (2012).
569. Chauvot de Beauchêne, I. *et al.* Hotspot Mutations in KIT Receptor Differentially Modulate Its Allosterically Coupled Conformational Dynamics: Impact on Activation and Drug Sensitivity. *PLoS Comput. Biol.* **10,** e1003749 (2014).
570. Da Silva Figueiredo Celestino Gomes, P. *et al.* Differential Effects of CSF-1R D802V and KIT D816V Homologous Mutations on Receptor Tertiary Structure and Allosteric Communication. *PLoS ONE* **9,** e97519 (2014).
571. Allain, A. *et al.* Allosteric pathway identification through network analysis: from molecular dynamics simulations to interactive 2D and 3D graphs. *Faraday Discuss* **169,** 303–321 (2014).
572. Penev, P. & Atick, J. Local feature analysis: a general statistical theory for object representation. *Netw. Comput. Neural Syst.* **7,** 477–500 (1996).
573. Zhang, Z. & Wriggers, W. Local feature analysis: a statistical theory for reproducible essential dynamics of large macromolecules. *Proteins* **64,** 391–403 (2006).

574. Zhang, Z. & Wriggers, W. Coarse-graining protein structures with local multivariate features from molecular dynamics. *J.Phys.Chem.B* **112,** 14026–14035 (2008).
575. Tchertanov, L., laine, E., Allain, A. & Chauvot de Beauchêne, I. MOdular NETwork Analysis (MONETA) version 2.0.
576. Wallace, A. C., Laskowski, R. A. & Thornton, J. M. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng.* **8,** 127–134 (1995).
577. Ihaka, R. & Gentleman, R. R: A Language for Data Analysis and Graphics. *J. Comput. Graph. Stat.* **5,** 299–314 (1996).
578. Mathieu Bastian, Sebastien Heymann & Mathieu Jacomy. Gephi: An Open Source Software for Exploring and Manipulating Networks. (2009).
579. Levitt, D. G. & Banaszak, L. J. POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J. Mol. Graph.* **10,** 229–234 (1992).
580. Laskowski, R. A. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.* **13,** 323–330, 307–308 (1995).
581. Peters, K. P., Fauck, J. & Frömmel, C. The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *J. Mol. Biol.* **256,** 201–213 (1996).
582. Liang, J., Woodward, C. & Edelsbrunner, H. Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. *Protein Sci.* **7,** 1884–1897 (1998).
583. Brady, G. P. & Stouten, P. F. Fast prediction and visualization of protein binding pockets with PASS. *J. Comput. Aided Mol. Des.* **14,** 383–401 (2000).
584. Weisel, M., Proschak, E. & Schneider, G. PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chem. Cent. J.* **1,** 7 (2007).
585. Le Guilloux, V., Schmidtke, P. & Tuffery, P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics* **10,** 168 (2009).
586. Durrant, J. D., de Oliveira, C. A. F. & McCammon, J. A. POVME: An algorithm for measuring binding-pocket volumes. *J. Mol. Graph. Model.* **29,** 773–776 (2011).
587. Laurent, B. *et al.* Epock: rapid analysis of protein pocket dynamics. *Bioinformatics* (2014). doi:10.1093/bioinformatics/btu822
588. Paramo, T., East, A., Garzón, D., Ulmschneider, M. B. & Bond, P. J. Efficient Characterization of Protein Cavities within Molecular Simulation Trajectories: *trj_cavity*. *J. Chem. Theory Comput.* **10,** 2151–2164 (2014).
589. An, J., Totrov, M. & Abagyan, R. Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol. Cell. Proteomics MCP* **4,** 752–761 (2005).
590. Laurie, A. T. R. & Jackson, R. M. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinforma. Oxf. Engl.* **21,** 1908–1916 (2005).
591. Schmidtke, P., Le Guilloux, V., Maupetit, J. & Tuffery, P. fpocket: online tools for protein ensemble pocket detection and tracking. *Nucleic Acids Res.* **38,** W582–W589 (2010).
592. Schmidtke, P., Bidon-Chanal, A., Luque, F. J. & Barril, X. MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories. *Bioinformatics.* **27,** 3276–3285 (2011).
593. Glaser, F. *et al.* ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinforma. Oxf. Engl.* **19,** 163–164 (2003).
594. Landau, M. *et al.* ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.* **33,** W299–302 (2005).
595. Ashkenazy, H., Erez, E., Martz, E., Pupko, T. & Ben-Tal, N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.* **38,** W529–W533 (2010).

596. Liu, B. A. *et al.* The human and mouse complement of SH2 domain proteins-establishing the boundaries of phosphotyrosine signaling. *Mol.Cell* **22,** 851–868 (2006).
597. Liu, B. A., Engelmann, B. W. & Nash, P. D. The language of SH2 domain interactions defines phosphotyrosine-mediated signal transduction. *FEBS Lett* **586,** 2597–2605 (2012).
598. Benes, C. H. *et al.* The C2 Domain of PKCδ Is a Phosphotyrosine Binding Domain. *Cell* **121,** 271–280 (2005).
599. Christofk, H. R., Vander Heiden, M. G., Wu, N., Asara, J. M. & Cantley, L. C. Pyruvate kinase M2 is a phosphotyrosine-binding protein. *Nature* **452,** 181–186 (2008).
600. Ren, J. *et al.* PhosSNP for Systematic Analysis of Genetic Polymorphisms That Influence Protein Phosphorylation. *Mol. Cell. Proteomics* **9,** 623–634 (2010).
601. Johnson, L. N. & Lewis, R. J. Structural basis for control by phosphorylation. *Chem. Rev.* **101,** 2209–2242 (2001).
602. Pawson, T. & Scott, J. D. Protein phosphorylation in signaling – 50 years and counting. *Trends Biochem. Sci.* **30,** 286–290 (2005).
603. Johnson, L. N. The regulation of protein phosphorylation. *Biochem. Soc. Trans.* **37,** 627 (2009).
604. Schlessinger, J. Cell signaling by receptor tyrosine kinases. *Cell* **103,** 211–225 (2000).
605. Olsen, J. V. *et al.* Global, In Vivo, and Site-Specific Phosphorylation Dynamics in Signaling Networks. *Cell* **127,** 635–648 (2006).
606. Chebaro, Y. *et al.* Phosphorylation of the Retinoic Acid Receptor Alpha Induces a Mechanical Allosteric Regulation and Changes in Internal Dynamics. *PLoS Comput. Biol.* **9,** e1003012 (2013).
607. De Groot, B. L., Hayward, S., van Aalten, D. M., Amadei, A. & Berendsen, H. J. Domain motions in bacteriophage T4 lysozyme: a comparison between molecular dynamics and crystallographic data. *Proteins* **31,** 116–127 (1998).
608. Skjaerven, L., Martinez, A. & Reuter, N. Principal component and normal mode analysis of proteins; a quantitative comparison using the GroEL subunit. *Proteins Struct. Funct. Bioinforma.* **79,** 232–243 (2011).
609. Berendsen, H. Collective protein dynamics in relation to function. *Curr. Opin. Struct. Biol.* **10,** 165–169 (2000).
610. Micheletti, C., Carloni, P. & Maritan, A. Accurate and efficient description of protein vibrational dynamics: Comparing molecular dynamics and Gaussian models. *Proteins Struct. Funct. Bioinforma.* **55,** 635–645 (2004).
611. Kiu, H. & Nicholson, S. E. Biology and significance of the JAK/STAT signalling pathways. *Growth Factors* **30,** 88–106 (2012).
612. Scaglia, P. A. *et al.* A Novel Missense Mutation in the SH2 Domain of the *STAT5B* Gene Results in a Transcriptionally Inactive STAT5b Associated with Severe IGF-I Deficiency, Immune Dysfunction, and Lack of Pulmonary Disease. *J. Clin. Endocrinol. Metab.* **97,** E830–E839 (2012).
613. Nelson, E. A., Sharma, S. V., Settleman, J. & Frank, D. A. A chemical biology approach to developing STAT inhibitors: molecular strategies for accelerating clinical translation. *Oncotarget.* **2,** 518–524 (2011).
614. Pinz, S. *et al.* The Synthetic α-Bromo-2′,3,4,4′-Tetramethoxychalcone (α-Br-TMC) Inhibits the JAK/STAT Signaling Pathway. *PLoS ONE* **9,** e90275 (2014).
615. Pinz, S., Unser, S. & Rascle, A. The Natural Chemopreventive Agent Sulforaphane Inhibits STAT5 Activity. *PLoS ONE* **9,** e99391 (2014).
616. Siggers, T. W. & Honig, B. Structure-based prediction of C2H2 zinc-finger binding specificity: sensitivity to docking geometry. *Nucleic Acids Res.* **35,** 1085–1097 (2007).
617. Rohs, R. *et al.* Origins of Specificity in Protein-DNA Recognition. *Annu. Rev. Biochem.* **79,** 233–269 (2010).

Annexes

Annexe A. Alignement des séquences primaires de la famille des protéines STATs humaines. Les résidus conservés sont sur fond coloré et les résidus similaires sont écrits en couleur.



Annexe B. Script de détection des ponts d'eau.

```
#!/usr/bin/perl

# WARNING: Development version !
# This script may NOT work or produce errors as it has been tested on perl 5.12 only.
#
# Hbridge.pl - Creates a xpm file and summaries the time occupancy of all Water Bridges (WB)
# between two molecules or selections (e.g. protein - DNA, protein - ligand ...) along a MD
        simulation.
# It also computes the molecules - water for each input xpm/ndx paired files.
#
# The hbmapPreProc, hbmapCoord and hbmapOutput subroutines are based on Justin
        Lemkul plot_hbmap.pl script,
# freely available at http://www.bevanlab.biochem.vt.edu/Pages/Personal/justin/scripts.html .
#
# This script needs as input files:
#   1. structure.pdb - a coordinate file (for atom naming), MUST be a .pdb file with NO CHAIN
        IDENTIFIERS !!!
#     As these file describes large systems ( < 9999 residues), residue numbers MUST be in
        hexadecimal!!!
#   2. mol1/2_water_map.xpm - 2 xpm files for molecules 1 and 2, as computed by gmx
#   3. mol1/2_water_map.ndx - 2 index files modified to contain only the atom numbers in the
        [hbonds...] section, NOTHING ELSE !!!
#     The atom number MUST NOT be in hexadecimal!!!
#
# If no output argument (sum1, sum2, map, sum) is provided, the default output files are:
# mol1_water.dat, mol2_water.dat, mol1_mol2.xpm and mol1_mol2.dat
#
# As proteins (or others macromolecules) can make a huge number of hydrogen bonds with
        water molecules
# along a MD trajectory, it is strongly advised to select only subsets of residues rather than
        entire macromolecules
# for the required g_hbond steps prior to this script. This will save memory and time.
#
# For those looking for internal bridges (i.e. mol1 = mol2), be aware that duplicates might exist.
# An additional function to deal with duplicates is under development.
#
# To do list:
#   - improve memory usage
#   - write a gmx xpm2ps-readable header for the mol1/2_water.xpm files (correct y-axis labels
        and information lines, ...)
#   - delete duplicates (or triplicates...), i.e. gather mol1 - waterX - mol2 and mol1 - waterY -
        mol2
#     in one line mol1 - mol2, and write corresponding (and correct if possible) dat/xpm files.
#   - use the built-in function substr to read pdb file --> no need to delete chain identifiers from
        pdb file anymore.
#   - help message to display
#   - add a progression bar for the hbmapCoor and MolMol functions.
#   - write additional comments about arguments/returns of subroutines
```

```
48  #
49
50  my $t0 = time();
51
52  use strict;
53  use warnings();
54
55  use List::Compare;
56  use List::MoreUtils qw(indexes);
57  use threads;
58  use threads::shared;
59  use Thread::Queue;
60
61  unless(@ARGV) {
62     die "Usage: perl $0 -s structure.pdb -xpm1 mol1_water_map.xpm -ndx1
63              mol1_water_index.ndx -xpm2 mol2_water_map.xpm -ndx2 mol2_water_map.ndx
64              -sum1 mol1_water_summary.dat -sum2 mol2_water_summary.dat
65              -map mol1_mol2_map.xpm -sum mol1_mol2_summary.dat -nt 4\n";
66  }
67
68  ################################################################################
69  # Store the scripts arguments in variables
70
71  # define input hash
72  my %args = @ARGV;
73
74  # input variables
75  my $struct;
76  my $mol1_map;
77  my $mol1_ndx;
78  my $mol2_map;
79  my $mol2_ndx;
80  my $max_cpus;
81
82  # catch up the input variables from the arguments
83  if (exists($args{"-s"})) {
84     $struct = $args{"-s"};
85  } else {
86     die "No -s specified!\n";
87  }
88
89  if (exists($args{"-xpm1"})) {
90     $mol1_map = $args{"-xpm1"};
91  } else {
92     die "No -xpm1 specified!\n";
93  }
94
95  if (exists($args{"-ndx1"})) {
96     $mol1_ndx = $args{"-ndx1"};
97  } else {
98     die "No -ndx1 specified!\n";
```

```
 99    }
100
101    if (exists($args{"-xpm2"})) {
102       $mol2_map = $args{"-xpm2"};
103    } else {
104       die "No -xpm2 specified!\n";
105    }
106
107    if (exists($args{"-ndx2"})) {
108       $mol2_ndx = $args{"-ndx2"};
109    } else {
110       die "No -ndx2 specified!\n";
111    }
112
113    if (exists($args{"-nt"})) {
114       $max_cpus = $args{"-nt"};
115    } else {
116       $max_cpus = 4;
117    }
118
119    # output variables
120    my $mol1_summary;
121    my $mol2_summary;
122    my $WBxpm;
123    my $WBdat;
124
125    # define output variable from the arguments
126    if (exists($args{"-sum1"})) {
127       $mol1_summary = $args{"-sum1"};
128    } else {
129       $mol1_summary = "mol1_water.dat";
130    }
131
132    if (exists($args{"-sum2"})) {
133       $mol2_summary = $args{"-sum2"};
134    } else {
135       $mol2_summary = "mol2_water.dat";
136    }
137
138    if (exists($args{"-sum"})) {
139       $WBdat = $args{"-sum"};
140    } else {
141       $WBdat = "mol1_mol2.dat";
142    }
143
144    if (exists($args{"-map"})) {
145       $WBxpm = $args{"-map"};
146    } else {
147       $WBxpm = "mol1_mol2.xpm";
148    }
149
```

```
150  ###############################################################################
151  ###
152  # MAIN Function of this script
153
154  # Preprocessing mol1 & mol2 data
155
156  print "\nPre-processing files $mol1_map and $mol1_ndx...\n";
157
158  my ($nhbonds1, $nframes1, $mol1_donors, $mol1_acceptors, $mol1_map, $Xheader)
159  = hbmapPreProc($mol1_map, $mol1_ndx);
160  my @mol1_header = @$Xheader;
161
162  print "\nPre-processing files $mol2_map and $mol2_ndx...\n";
163
164  my ($nhbonds2, $nframes2, $mol2_donors, $mol2_acceptors, $mol2_map, $Xheader)
165              = hbmapPreProc($mol2_map, $mol2_ndx);
166  my @mol2_header = @$Xheader;
167
168  my @mol1_donor_resName    :shared;
169  my @mol1_donor_atomName   :shared;
170  my @mol2_donor_resName    :shared;
171  my @mol2_donor_atomName   :shared;
172  my @mol1_acceptor_resName   :shared;
173  my @mol1_acceptor_atomName  :shared;
174  my @mol2_acceptor_resName   :shared;
175  my @mol2_acceptor_atomName  :shared;
176
177  # open the structure pdb file and extract residue & atom names of each H-bond
178  open(STRUCT, "<$struct") || die "Cannot open input coordinate file!\n";
179  print "\nProcessing coordinate file for Mol-Water h-bonds atom names...\n";
180
181  my $coordinate = Thread::Queue->new();
182  my @threads;
183
184  foreach ( 1 .. $max_cpus) {
185     push( @threads, async { hbmapCoord($coordinate) } );
186  }
187
188  while(<STRUCT>) {
189     $coordinate->enqueue( $_ );
190  }
191
192  foreach my $thread ( @threads ) {
193     $coordinate->enqueue( undef );
194  }
195
196  foreach my $thread ( @threads ) {
197     $thread->join();
198  }
199  close(STRUCT);
200
```

```
# Write hbmap output files for mol1 and mol2
hbmapOutput($nhbonds1, \@$mol1_map, $mol1_summary, \@mol1_donor_resName,
        \@mol1_donor_atomName, \@mol1_acceptor_resName,
        \@mol1_acceptor_atomName, $nframes1);
hbmapOutput($nhbonds2, \@$mol2_map, $mol2_summary, \@mol2_donor_resName,
        \@mol2_donor_atomName, \@mol2_acceptor_resName,
        \@mol2_acceptor_atomName, $nframes2);

# By now, we have written 2 summaries for mol1_water and mol2_water H bonds.
# Let's begin the hard stuff: i.e. extracting water-bridges

# Control that data come from a similar simulation (i.e. same number of frames)
die "mol1_water and mol2_water xpm files have different number of frames!\nPlease provide
        input files with the same number of frames." if ($nframes1 != $nframes2);

# Write down final output files headers
print "Write final .xpm and .dat headers...\n";
open(XPM, ">>$WBxpm") || die "Cannot open water-bridge map file $WBxpm!\n";
printf(XPM "/* XPM */\n/* This file can be converted to EPS by the GROMACS program xpm2ps
        */\n/* title:\t\"Hydrogen Bond Existence Map\" */\n /* legend:\t\"Hydrogen Bonds\"
        */\n/* x-label:\t\"Time (ps)\" */\n/* y-label:\t\"Hydrogen Bond Index \" */\n/*
        type:\t\"Discrete\" */\nstatic char *gromacs_xpm[] = {\n"); # This line may be changed if
        the time is in ns rather than in ps.
printf(XPM "\"%3s %2s   2 1\"\n", $nframes1, ($nhbonds1 + $nhbonds1));   # Second value
        incorrect!! Should be the number of WB computed...
printf(XPM "\"   c #FFFFFF \" /* \"None\" */,\n\"o  c #FF0000 \" /* \"Present\" */,\n");
printf(XPM "@mol1_header");
#To be added: y-axis lines... The problem is we need to know how many Water-briges do exist
before calculating them. A workaround has to be found.

open(DAT, ">>$WBdat") || die "Cannot open water-bridge summary file $WBdat!\n";
printf(DAT "#    Molecule 1 \t Water \t Molecule 2 \t %% Exist.\n");

# Extract Water bridges
print "Processing the Water bridges...\n";

# Molmol threaded
print "\nProcessing the case Mol1...H-O-H...Mol2 with threads...\n";

my $index = Thread::Queue->new();
my @threads;

foreach ( 1 .. $max_cpus) {
   push( @threads, async { MolMol($index, \@mol1_donor_resName, \@mol2_donor_resName,
        \@$mol1_map, \@$mol2_map, \@mol1_acceptor_resName,
        \@mol1_acceptor_atomName, \@mol2_acceptor_resName,
        \@mol2_acceptor_atomName, $nframes1) } );
}

for (my $p = 1; $p < scalar(@mol1_donor_resName); $p++) {
   $index->enqueue( $p );
```

```
252     }
253
254     foreach my $thread ( @threads ) {
255        $index->enqueue( undef );
256     }
257
258     foreach my $thread ( @threads ) {
259        $thread->join();
260     }
261
262     printf("\nExploring the case Mol1...H-O...H-Mol2:\n");
263     my $index = Thread::Queue->new();
264     my @threads;
265
266     foreach ( 1 .. $max_cpus) {
267        push( @threads, async { MolMol($index, \@mol1_donor_resName,
268              \@mol2_acceptor_resName, \@$mol1_map, \@$mol2_map,
269              \@mol1_acceptor_resName, \@mol1_acceptor_atomName, \@mol2_donor_resName,
270              \@mol2_donor_atomName, $nframes1) } );
271     }
272
273     for (my $p = 1; $p < scalar(@mol1_donor_resName); $p++) {
274        $index->enqueue( $p );
275     }
276
277     foreach my $thread ( @threads ) {
278        $index->enqueue( undef );
279     }
280
281     foreach my $thread ( @threads ) {
282        $thread->join();
283     }
284
285     printf("\nExploring the case Mol1-H...O-H...Mol2:\n");
286     my $index = Thread::Queue->new();
287     my @threads;
288
289     foreach ( 1 .. $max_cpus) {
290        push( @threads, async { MolMol($index, \@mol1_acceptor_resName,
291              \@mol2_donor_resName, \@$mol1_map, \@$mol2_map, \@mol1_donor_resName,
292              \@mol1_donor_atomName, \@mol2_acceptor_resName,
293              \@mol2_acceptor_atomName, $nframes1) } );
294     }
295
296     for (my $p = 1; $p < scalar(@mol1_acceptor_resName); $p++) {
297        $index->enqueue( $p );
298     }
299
300     foreach my $thread ( @threads ) {
301        $index->enqueue( undef );
302     }
```

```
303
304     foreach my $thread ( @threads ) {
305       $thread->join();
306     }
307
308     printf("\nExploring the case Mol1-H...O...H-Mol2:\n");
309
310     my $index = Thread::Queue->new();
311     my @threads;
312
313     foreach ( 1 .. $max_cpus) {
314       push( @threads, async { MolMol($index, \@mol1_acceptor_resName,
315             \@mol2_acceptor_resName, \@$mol1_map, \@$mol2_map, \@mol1_donor_resName,
316             \@mol1_donor_atomName, \@mol2_donor_resName, \@mol2_donor_atomName,
317             $nframes1) } );
318     }
319
320     for (my $p = 1; $p < scalar(@mol1_acceptor_resName); $p++) {
321       $index->enqueue( $p );
322     }
323
324     foreach my $thread ( @threads ) {
325       $index->enqueue( undef );
326     }
327
328     foreach my $thread ( @threads ) {
329       $thread->join();
330     }
331
332     close(DAT);
333     close(XPM);
334
335     # To add here: function to delete duplicates
336             (mol1-W1203-mol2 + mol2-W1203-mol1 == mol1 - mol2)
337     my $t1 = time();
338     my $elapsed = $t1-$t0;
339     print "\nTotal time elapsed: $elapsed s\n\nEND\n\n";
340
341     exit;
342
343     # End of MAIN
344     ################################################################################
345
346     ############################ SUBROUTINE hbmapPreProc ############################
347
348     sub hbmapPreProc {
349
350       open(MAP, "<$_[0]") || die "Cannot open $_[0] input map file!\n";
351       my @xpm = <MAP>;
352       close(MAP);
353
```

```
354     open(NDX, "<$_[1]") || die "Cannot open $_[1] input index file!\n";
355     my @index = reverse <NDX>;  # Read the .ndx file top-down to fit xpm ordering
356     close(NDX);
357
358   # determine number of HB indices and frames for molecule-water interactions
359     my $nhbonds = 0;
360     my $nframes = 0;
361
362     for (my $i = 0; $i < 20; $i++) {
363
364       if ($xpm[$i] =~ 'static char') {
365
366         my $hbond_line = $xpm[$i+1];
367         my @info = split(" ", $hbond_line);
368
369         $nframes = $info[0];
370         my @nframes = split('', $nframes);
371         shift(@nframes);    # get rid of the "
372         $nframes = join('', @nframes);
373
374         $nhbonds = $info[1];
375         last;
376
377       }
378
379     }
380
381     print "From the $_[0] file, there are $nhbonds H-bond indices and $nframes frames.\n";
382
383   # clean up the header - the top 12 lines of comments, excluding the x-axis and y-axis lines
384     splice(@xpm, 0, 12);
385
386   # store "x-axis" or "y-axis" lines in new variables
387     my @Xheader = ();
388     my @Yheader = ();
389     my @map = ();
390
391     while (my $mapLine = shift(@xpm)) {
392       if ($mapLine =~ /x-axis/){
393         push(@Xheader, $mapLine);
394       } elsif ($mapLine =~ /y-axis/){
395         # Do nothing
396       } else {
397         push(@map, $mapLine);
398       }
399     }
400
401   # initialize donor/acceptor hashes for bookkeeping purposes     my %donors;
402   for (my $b=1; $b<=$nhbonds; $b++) {
403       $donors{$b} = 0;
404   }
```

```
405
406     my %acceptors;
407     for (my $c=1; $c<=$nhbonds; $c++) {
408        $acceptors{$c} = 0;
409     }
410
411  # Open the index files and put donnor's and acceptor's atom numbers in hashes
412     for (my $n=0; $n<$nhbonds; $n++) {
413        my @line = split(" ", $index[$n]);
414        $donors{$n+1} = $line[0];
415        $acceptors{$n+1} = $line[2];
416     }
417
418  # These three last steps will be gathered to save time.
419
420     return($nhbonds, $nframes, \%donors, \%acceptors, \@map, \@Xheader);
421
422  }
423
424  ############################ SUBROUTINE hbmapCoord ############################
425
426  sub hbmapCoord {
427
428     my( $q ) = @_;
429     while( defined( my $data = $q->dequeue() ) ) {
430
431        my @line = split(" ", $data);
432        if ($line[0] eq 'ATOM') {
433
434           while (my ($z, $atom) = each %$mol1_donors) {
435              if ( $atom eq ( hex $line[1] ) ) {
436                 $mol1_donor_atomName[$z] = $line[2];
437                 $mol1_donor_resName[$z] = join('', $line[3], $line[4]);
438              }
439           }
440
441           while (my ($z, $atom) = each %$mol1_acceptors) {
442              if ( $atom eq ( hex $line[1] ) ) {
443                 $mol1_acceptor_atomName[$z] = $line[2];
444                 $mol1_acceptor_resName[$z] = join('', $line[3], $line[4]);
445              }
446           }
447
448           while (my ($z, $atom) = each %$mol2_donors) {
449              if ( $atom eq ( hex $line[1] ) ) {
450                 $mol2_donor_atomName[$z] = $line[2];
451                 $mol2_donor_resName[$z] = join('', $line[3], $line[4]);
452              }
453           }
454
455           while (my ($z, $atom) = each %$mol2_acceptors) {
```

```
456                 if ( $atom eq ( hex $line[1] ) ) {
457                     $mol2_acceptor_atomName[$z] = $line[2];
458                     $mol2_acceptor_resName[$z] = join('', $line[3], $line[4]);
459                 }
460             }
461         }
462     }
463 }
464
465 ############################ SUBROUTINE hbmapOutput ############################
466
467 sub hbmapOutput {
468
469     my %hbonds;
470     for (my $a = 0; $a < $_[0]; $a++) {
471         $hbonds{$a+1} = 0;
472     }
473
474     my @map = @{$_[1]};
475     for (my $b = 1; $b <= $_[0]; $b++) {
476         $hbonds{$b} = grep(/o/, (split('', $map[$b-1])));
477     }
478
479 # Open, write and close the output file
480     open(OUT, ">>$_[2]") || die "Cannot open output file $_[2]!\n";
481     print "\nWriting output file $_[2]...\n";
482     printf(OUT "#    Donor \t\t\t Acceptor \t\t\t %% Exist.");
483
484     my @donor_resName = @{$_[3]};
485     my @donor_atomName = @{$_[4]};
486     my @acceptor_resName = @{$_[5]};
487     my @acceptor_atomName = @{$_[6]};
488
489     for (my $c = 1; $c <= $_[0]; $c++) {
490         printf(OUT "\n%10s\t%10s\t%10s\t%10s\t%10.3f", $donor_resName[$c],
491                             $donor_atomName[$c], $acceptor_resName[$c], $acceptor_atomName[$c],
492                             ( ( $hbonds{$c} / $_[7] ) * 100 ) );
493     }
494
495     close(OUT);
496
497 }
498
499 ############################## SUBROUTINE MolMol ##############################
500 # The MolMol function handles the hbmap outputs to extract Molecule1 - Molecule2 bridges
501 mediated by water molecules.
502
503 sub MolMol {
504     my @resName1 = @{$_[1]};
505     my @resName2 = @{$_[2]};
506     my @map1 = @{$_[3]};
```

```
507       my @map2 = @{$_[4]};
508       my @resName3 = @{$_[5]};
509       my @atomName3 = @{$_[6]};
510       my @resName4 = @{$_[7]};
511       my @atomName4 = @{$_[8]};
512       my $frames = $_[9];
513
514       my( $x ) = @_;
515       while( defined( my $p = $x ->dequeue() ) ) {
516
517          if (@resName1[$p] =~ m/SOL\w{1,6}/ ) {
518
519             for (my $q = 1; $q < scalar(@resName2); $q++) {
520                if (@resName2[$q] =~ @resName1[$p]) {
521                   my @mol1_indexes = indexes { $_ =~ /o/} (split('', @map1[$p-1]));
522                   my @mol2_indexes = indexes { $_ =~ /o/} (split('', @map2[$q-1]));
523                   my $indexes_comp = List::Compare->new(\@mol1_indexes, \@mol2_indexes);
524                   my @intersect = $indexes_comp->get_intersection;
525                   my $new_counter = $#intersect + 1;
526
527                   if ($new_counter != 0 ) {
528
529                      my $templine;
530
531                      foreach my $framendx (1..$frames) {
532
533                         if ($framendx ~~ @intersect){
534                            $templine = $templine . "o";
535                         }
536                         else {
537                            $templine = $templine . " ";
538                         }
539
540                         $framendx++;
541                      }
542
543                      printf(XPM "\"$templine\",\n");
544                      printf(DAT "%10s\t%10s\t%10s\t%10s\t%10s\t%10.3f\n", @resName3[$p],
545                                 @atomName3[$p], @resName1[$p], @resName4[$q], @atomName4[$q],
546                                 (($new_counter / $nframes1) * 100));
547                      # Should return the line rather than printing it,  # as it allows simultaneus printing
548                                 from different threads!!
549                   }
550                }
551             }
552          }
553       }
554    }
555
556    ########################## END ###############################################
```